Санкт-Петербургский Государственный Университет

Кафедра Математического Моделирования Энергетических Систем

Васин Алексей Геннадиевич

Выпускная квалификационная работа бакалавра

Вероятностно-статистический анализ динамики успеваемости в ВУЗе

Направление 010400

Прикладная математика фундаментальная информатика и программирование

Научный руководитель,

кандидат физ.-мат. наук,

доцент

Свиркин М.В.

Санкт-Петербург

2016

Содержание

Введение

Образование - одна из важнейших составляющих жизни социума. В современном мире специалисты высокого уровня играют особую роль. Один из ключевых факторов становления подобного рода специалистов это высшее образование. Высшее образование зачастую играет ключевую роль, как в индивидуальной, так и в коллективной эффективности. Поэтому сегодня как никогда важно иметь, как можно более точную оценку качества образования. Один из самых распространённых критериев которого - оценка знаний студента экспертами. Имеется в виду самая распространённая оценка от преподавателя в ВУЗе. Отметки успеваемости универсантов назначаются в соответствии с их прогрессом в учёбе: сдаче контрольных работ, экзаменов, зачётов, рефератов и т.п. Подобные баллы являются естественным и универсальным показателем, позволяющим выявить степень усвоения материала; также на основании оценок можно определить тенденции и сущности, влияющие на успеваемость. Для проведения подобных исследований широко применяются методы теории вероятности и математической статистики, к примеру, корреляционный-регрессионный анализ. В таком анализе, помимо стандартных методов статистики, следует учитывать временную динамику успеваемости, что позволит более полно раскрыть картину качества образования. Глубокий вероятностно-статистический анализ данных, позволяет выявить ключевые тенденции в успешности образования. Что впоследствии даёт возможность влияния на качество учебного процесса принятием оптимальных управленческих решений.

Постановка задачи

В данной работе рассматривается выборка из данных об успеваемости студентов бакалавров и специалистов факультета ПМ-ПУ различных годов поступления по соответствующим предметам и соответствующим годам. Данные представлены таблицами в формате .xls, где по строкам закодированные номера студентов, а по строкам - различная информация. Например, такая как успеваемость, баллы ЕГЭ и т.п. данные. Каждая таблица соответствует выборке студентов одного года поступления в определённую сессию, одной или нескольких укрупнённых групп. Пример таблицы данных приводится в разделе приложений.

Данные				
Учебный год	**Год поступления**	**Зимняя/летняя сессия**	**Курс**	**Объём выборки**
2010/2011				
	2010	Зима	1	234
2011/2012				
	2009	Зима	3	174
	2010	Зима	2	189
	2011	Зима	1	270
2012/2013				
	2009	Лето	4	160
	2010	Лето	3	152
	2011	Лето	2	179
	2012	Лето	1	225
	2009	Зима	4	310
	2010	Зима	3	273
	2011	Зима	2	271
	2012	Зима	1	287

2013/2014				
	2009	Зима	5	310
	2010	Зима	4	273
	2011	Зима	3	271
	2012	Зима	2	287
	2013	Зима	1	280

Данные предоставлены управлением службы информационных технологий ФГБОУ ВО "Санкт Петербургский государственный университет".

Нарушений положения о персональной информации нет, в силу того, что личные данные нигде не используются в явном виде.

Целью данной научной работы является исследование успеваемости студентов, с учётом временной динамики данных, с помощью вероятностно-статистического аппарата.

Обзор литературы

Проблемам статистического анализа успеваемости посвящена научная литература. Данный параграф посвящён обзору статей, и рассмотренных в них вопросах, прямым и непосредственным образом связанных с настоящей работой.

Так в работах [2,11] подтверждено, что *распределение средних сессионных оценок по математическим дисциплинам, при устоявшемся учебном процессе, может быть описано законом нормального распределения*. Рекомендовано, в частности, вести не только множественный, но и погруповой анализ динамики факторов, а также проводить психологические исследования и мотивационные курсы для универсантов.

Но средний бал зачётки, по существу, не учитывает многих факторов. Таких, например, как пересдача на более высокую оценку, отчисление студента или общее количество, не допущенных до экзамена студентов. Многие психофизиологические факторы, влияющие на успеваемость, исследованы в работах [6, 9]. Также ответы на вопросы даёт работа [3].В качестве показателей успеваемости в работе используются традиционные средние оценки коллектива студентов, но наряду с ними вводятся скорректированные средние оценки. *Скорректированные средние оценки, в отличие от традиционных, учитывают число не допущенных к экзамену и отчисленных от обучения в вузе студентов. На исследуемых этапах скорректированная средняя оценка коллектива студентов заметно изменяется, в то время как традиционная средняя оценка сохраняет*

постоянное значение. Из чего в статье следует закономерный вывод, что скорректированные средние оценки оказываются более содержательными, чем традиционные оценки, и позволяют более четко и адекватно проследить динамику уровня успеваемости коллектива студентов. Поэтому статистический анализ, как заключено в данной статье, динамики изменения успеваемости студентов рекомендуется проводить с учетом скорректированных показателей.

Но способ подсчёта средних оценок не единственное на что следует обратить внимание. Так, к примеру, переход от непараметрической статистики успеваемости отдельного студента к параметрической статистике успеваемости большой (> 100 человек) группы студентов оказывается не вполне тривиальным [4]. *Ограниченность диапазона допустимых значений средней успеваемости при приближении к границе приводит к возникновению зависимости между средней дисперсией в локальном интервале и средней успеваемостью в том же интервале*. В вероятностно-статистическом моделировании обучающего процесса зачастую используется нормальное распределение, которое обычно успешно проходит проверку на согласованность. Однако ряд исследований [4,7] показывают что, не смотря на то, что значение критерия, к примеру, Хи-квадрат, не отвергает гипотезу о нормальности полученного распределения, *можно утверждать, что выборка не будет подчиняться распределению Гаусса в связи с наличием зависимости [7] дисперсии от значения средней успеваемости студентов. Также результаты пересдач неудовлетворительных оценок искажают статистическую картину успеваемости.* Как следствие, нецелесообразно применение предельного перехода от непараметрической статистики отдельных студентов к

параметрической статистике общей выборки студентов без предварительного выделения подгруппы студентов, отвечающих условиям применимости предельной теоремы.

В связи с этим эффектом средние успеваемости, найденные для больших выборок студентов, могут не подчиняться нормальному закону распределения случайной величины. Естественным способом преодоления выявленных проблем является использование вероятностных характеристик [4] успеваемости, которые одинаковым способом рассчитываются как для отдельного студента, так и для большой выборки студентов. Вероятность получения студентом определённой оценки призвана не заменить, но дополнить и раскрыть характеристику средней успеваемости, предоставить более информативные данные для сопоставления с результатами педагогических исследований.

Иными словами, в работе подтвердилось наличие зависимости между средней успеваемостью и вероятностными характеристиками получения каждой из возможных оценок студентами использованной выборки с помощью методов регрессионного анализа. *Частным результатом использования вероятностных характеристик успеваемости является то, что оценка «хорошо» является компромиссной. Она имеет две фундаментальные тенденции: четыре ближе к трём, и четыре ближе к пяти.* Из статьи также следует существование двух ветвей графика успеваемости студентов, поэтому можно предполагать наличие различий и в факторном пространстве, описывающем мотивацию к обучению. Для подобной вероятностной методики снимается проблема предельного перехода, а также она позволяет сделать предположение о том, что

вероятностные характеристики успеваемости студентов являются достаточно информативными.

Кроме описательной статистики в анализе успеваемости применяются также методы корреляционно-регрессионного анализа [1, 4, 5, 7, 8, 10]. Так одна из статей [1] получает косвенный вывод о том, что аппарат линейной регрессии, даже с учётом зависимости, проверенной через множественный коэффициент детерминации, данных друг от друга, не всегда адекватно справляется с поставленной задачей предсказания успеваемости по специальному предмету в зависимости от успеваемости по другим предметам. Вследствие этого следует тщательным образом проводить проверку на всех этапах исследования.

Немаловажным фактором в успешности обучения являются и исходные факторы, к примеру, экспертная оценка знаний абитуриентов, иными словами баллы ЕГЭ, на основании которых может анализироваться успешность обучения в вузе [5, 8, 12]. В частности, не подтверждается *гипотеза о том, что баллы ЕГЭ по разным предметам в равной степени эффективны в качестве прогностического фактора дальнейшей успеваемости.* С помощью методов регрессионного анализа оценивается сила связи результатов вступительных экзаменов (как суммарного балла ЕГЭ, так и баллов по отдельным предметам) и успеваемости студентов в вузе. *В среднем, по результатам перечисленных исследований, баллы ЕГЭ объясняют 20–30% вариации успеваемости студентов в вузе. Стандартизированные вступительные экзамены широко используются в мировой практике, к примеру, это — SAT и ACT в США и Matura в ряде европейских стран. Основной подход к анализу валидности SAT и ACT — оценка силы линейной взаимосвязи между результатами тестов*

и показателями успеваемости студентов в вузе с помощью расчета коэффициента корреляции Пирсона или регрессионных моделей, где зависимая переменная — показатель успеваемости, а предикторы — баллы SAT или ACT. В роли итогового показателя качества предсказания выступают величина коэффициента корреляции, возведенная в квадрат, или коэффициент детерминации в регрессионных моделях, интерпретируемые как доля вариации зависимой переменной, которая объясняется независимыми переменными. *Величины соответствующих коэффициентов для SAT и ACT находятся в интервале от 0,35 до 0,46 с учетом стандартной ошибки. Таким образом, вступительные экзамены предсказывают 12–25% вариации оценок в вузе (величина R^2). В ряде работ наряду с результатами вступительных испытаний SAT или ACT учитывается средняя оценка школьного аттестата. Она зачастую оказывается лучшим предиктором успеваемости в вузах, чем баллы SAT или ACT, а совместный учет результатов вступительных экзаменов и средней школьной оценки в одной модели значительно повышает успешность предсказания академических достижений.*

Ссылаясь на многочисленные исследования, в статье говорится, что именно 1-й год учебы в вузе является определяющим для успеваемости на всех последующих курсах, сдачи итоговых экзаменов, и даже для успеваемости в магистратуре. Поэтому важным условием прогностической валидности экзамена является его способность предсказывать успеваемость именно на 1-м курсе.

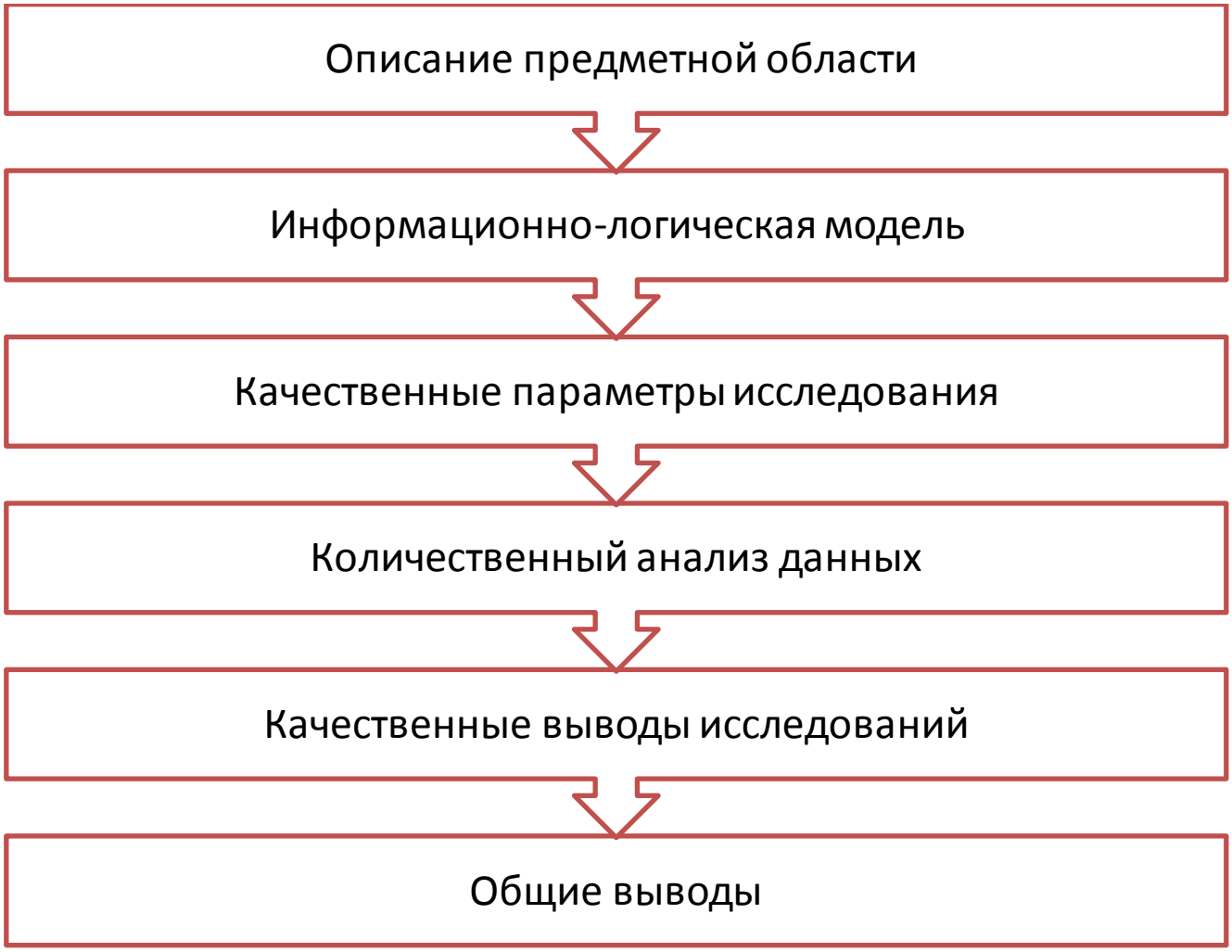
В работе проводится регрессионный анализ, в соответствии с общепринятой методологией, измеряется связь результатов ЕГЭ и дальнейшей успеваемости в вузе с помощью линейного регрессионного

анализа. В качестве вывода в статье говорится о том, что баллы ЕГЭ в наибольшей степени влияют на успеваемость в университете на 1-м году учебы. Установлено, что результаты ЕГЭ связаны с успеваемостью на 2м курсе и на старших курсах только опосредованно — через успеваемость на 1-м курсе. Из чего следует, что в дальнейшем следует рассматривать только связь ЕГЭ с успеваемостью на 1-м курсе, а её способность предсказывать итоги 1-го года учебы можно считать достаточной для того, чтобы считать экзамен валидным и пригодным в качестве критерия.

Предсказательная способность баллов ЕГЭ по отдельным предметам, составляющих суммарный итоговый балл ЕГЭ, примерно одинакова, но все же ЕГЭ по математике и русскому языку являются лучшими предикторами для подавляющего большинства направлений. При сравнении групп студентов, поступивших в вуз на основании результатов ЕГЭ и как победители олимпиад, выясняется, что разница между ними как в среднем балле ЕГЭ, так и в средней оценке за 1-й курс имеет место только на тех факультетах, где отбор преимущественно идет по математическим и экономическим олимпиадам.

Глава 1. Информационно-логическая модель успеваемости

1.1 Структура исследования



1.2 Описание и формализация предметной области

Одна из важнейших оценок успеваемости студента - это мнение эксперта о сформированной компетенции учащегося в заданной сфере знаний. Оценка имеющихся навыков и эрудиции производится различными

способами. Это могут быть тестовые вопросы, проверка навыков на практической задаче, написание курсовых и реферативных заданий, доклад по соответствующему предмету, беседа с преподавателем на заданную тему или другие способы оценить знания универсанта. Также следует учитывать различные факторы, которые могут повлиять на оценку успеваемости студента.

Немаловажную роль в успеваемости универсанта играет школьная подготовка, которая в свою очередь оценивается с помощью единого государственного экзамена, призванного дать независимую и наиболее полную оценку сформированного багажа знаний абитуриента на момент поступления в ВУЗ. Такие предметы как математика, русский язык, информатика и физика являются наиболее значимыми для поступления на факультет ПМ-ПУ. Как следствие именно по баллам этих экзаменов и проводится отбор среди абитуриентов. Конечно, проходной балл из года в год изменяется, в связи и с изменением сложности ЕГЭ, и с изменением конкурсного отбора.

Но оценка знаний поступающего может быть произведена не только с помощью единого экзамена, абитуриент также может подтвердить свою компетенцию путём занятия призового места на олимпиадах определённого уровня. После чего либо получит максимальный балл по соответствующему предмету, либо сразу будет зачислен на факультет. Это происходит, когда профиль олимпиады совпадает с профилем обучения. Как показывают исследования, призёры олимпиад оказывают достойную конкуренцию другим абитуриентам.

Различные города имеют свою уникальную культуру и традиции. Также следует учитывать общий интеллектуальный климат в регионе. Не менее важен регион, откуда поступает абитуриент, и для психо-эмоционального состояния студента. Учится ли универсант в родном регионе или нет.

Естественно следует отметить и другие особые обстоятельства зачисления в университет. К таким обстоятельствам может относиться принадлежность абитуриента к группам людей, которым государство уделяет особое внимание. Это, например, такие люди как инвалиды и сироты. Тенденции в успеваемости этих групп людей могут отличаться от остальных.

Как известно, в ВУЗах есть как бюджетные, так и коммерческие места. Учёба на коммерческих местах связана с финансовыми затратами, а также с отдельным конкурсом. Что часто упрощает поступление.

Кроме коммерческих мест существует также вечернее отделение. Государство позаботилось о тех, кто, несмотря на работу, хочет получить высшее образование. Такие люди, естественно, имеют меньше свободного времени, что сказывается на успешности усвоения материала.

Целью университета является подготовка конкурентно способного специалиста. Профессионала, который впоследствии будет приносить пользу обществу. В связи, с чем инициатива работодателей по подготовке высококлассного специалиста, путём обучения его в ВУЗе, приобретает определённый характер. Студенты, зачисленные по целевому набору, могут иметь отличную от других мотивацию к учёбе, а также иные возможности. Что в определённой мере влияет на их прогресс в учёбе.

Другими факторами являются предметы, по которым выставляется отметка. Предметы могут быть как гуманитарными, к примеру, английский, история Санкт-Петербурга или философия, так и точными, а это: алгебра, дифференциальные уравнения, геометрия и т.п.

Также следует учитывать разделение на предметы связанные с программированием, например архитектуры вычислительных систем, технология программирования, изучение различных языков программирования и пакетов прикладных программ, и связанные с классическим математическим образованием, такие как математический анализ, уравнения в частных производных и теория вероятности.

Направление подготовки студента задаёт лишь общий характер предметов. Для более узкой и гибкой специализации студентов действует система курсов по выбору. Эта система помогает получить знания соответствующие талантам и наклонностям студента, что обеспечивает как более высокий профессионализм, так и лучшую психологическую обстановку для успешного обучения.

Подготовка профессионалов во всех сферах математики является неэффективной задачей, с точки зрения соотношения затрат и эффективности. Поэтому в ВУЗах существуют различные направления подготовки специалистов. Такие, например, как “Прикладная математика и информатика”, “Прикладная математика и физика” или “Информационные технологии”. В связи с различными компетентностями успеваемость студентов разных направлений может отличаться.

1.3 Информационно-логическая модель успеваемости

Обучение в ВУЗе	Факторы поступления	По ЕГЭ (Математиа, Информатика, Физика и Русский язык)
		По Олимпиаде
		Целевой набор
		Регион поступления
		Особые обстоятельства
	Форма обучения	Направление обучения
		Дневное/вечернее отделение
		Коммерческая/бюджетная форма обучения
		Восстановлен
	Предметы	Математические или гуманитарные
		Информационные технологии или чистая математика
		Общие или специальные
		Обязательные, курсы по выбору или факультативы
	Временная динамика	Учебный год
		Зимняя/летняя сессии
		Год поступления

1.4 Задачи

Из информационно-логической модели естественным образом следуют направления исследования:

- Успеваемость и поступление
- Успеваемость, в разрезе различных форм обучения
- Успеваемость по предметам
- Успеваемость во временной динамике

Кроме того, следует учесть ряд вопросов, возникающих из обзора литературы, а также из теории вероятности и математической статистики:

- Исследование зависимости между дисперсией и средним баллом за экзамен
- ЕГЭ и успеваемость на различных курсах
- Общий анализ факторов, влияющих на успеваемость
- Исследование согласия успеваемости с нормальным законом распределения

Вышеизложенные соображения реализуют себя в следующих задачах, решаемых в данной работе:

1. Многофакторный анализ успеваемости
2. Исследование связи баллов ЕГЭ и отметок в контексте временной динамики
3. Изучение успеваемости в контексте предметов
4. Проверка согласия успеваемости с нормальным законом распределения

5. Исследование зависимости между дисперсией и средним баллом за экзамен, проверка на подобную зависимость взаимосвязи дисперсии и среднеквадратичного отклонения у среднего балла студента в сессию
6. Исследование среднеквадратичных отклонений отметок за экзамен
7. Сравнение успеваемости зимних и летних сессий

1.5 Первичная обработка и организация данных

Начальные данные представлены xls таблицами. Каждая таблица – успеваемость студентов одного курса в определённую сессию, а также некоторая информация о самих студентах. По строкам таблиц – закодированные Ф.И.О. студента. По столбцам таблицы – данные о различных факторах, относящихся к процессу обучения студента в ВУЗе. Пример таблицы приведён в разделе приложений.

С целью упрощения работы с данными было проведено разделение, унификация и кодировка данных. Каждый новый файл также представляет собой таблицу, организованную по принципу, описанному выше. Была проведена кодировка самих файлов. Пример кодировки файла - Zima2013x2014Yp2013p010000. Что значит, зимняя сессия 2013-2014 года студентов 2013 года поступления, номер укрупнённой группы студентов - 010000.

Для удобства оперирования и исследования массивов данных, была организована сводная таблица DescrStat. При организации таблицы, были автоматически подсчитаны некоторые показатели. Пример таблицы приведён в разделе приложений.

Организация сводной таблицы:

Информация	**Имя столбца, содержащего необходимую информацию**
Имя файла	File

Год поступления	Yp
Учебный год	Y
Маркировка зимней или летней сессии	ZimaLeto
Направление обучения	Napr
Курс	kurs
Размер выборки	Size
Числовая кодировка схожих экзаменов	sr
Маркировка информационных, математических или гуманитарных предметов	ItMathHum
Обязательные и необязательные предметы (маркировка)	ObyazViF
Наименование экзамена. В дополнение - средний балл студента в сессию	Parameter
Средний балл за экзамен	nanmean
Среднеквадратичное отклонение баллов за экзамен	nanstd
Медиана	nanmedian
Оценка коэффициента асимметрии	skewness
Разность между 75% и 25%	iqr

квантилями	
Оценка коэффициента эксцесса	kurtosis
Далее следуют p-значения для коэффициентов корреляции и коэффициенты корреляции	
Средний балл и оценка за успеваемость (p-значение)	CorrPMrSr
Средний балл и оценка за успеваемость (корреляция)	CorrRMrSr
ЕГЭ по математике и оценка успеваемости (p-значение)	CorrPMrMath
ЕГЭ по математике и оценка успеваемости (корреляция)	CorrRMrMath
ЕГЭ по информатике и оценка успеваемости (p-значение)	CorrPMrInf
ЕГЭ по информатике и оценка успеваемости (корреляция)	CorrRMrInf
ЕГЭ по физике и оценка успеваемости (p-значение)	CorrPMrPhys
ЕГЭ по физике и оценка успеваемости (корреляция)	CorrRMrPhys
ЕГЭ по русскому и оценка успеваемости (p-значение)	CorrPMrRus
ЕГЭ по русскому и оценка успеваемости (корреляция)	CorrRMrRus

Глава 2. Математические методы

Критерий Ярке—Бера — статистический критерий, для проверки согласия распределения с нормальным законом, посредством сверки его третьего и четвёртого моментов с соответствующими моментами нормального распределения.

В критерии выдвигается гипотеза $H_0: K = 0, S = 3$ против гипотезы $H_0: K \neq 0, S \neq 3$, где S — коэффициент асимметрии, K — коэффициент эксцесса.

При значительном числе наблюдений *n*, статистика

$$JB = \frac{n}{6}(s^2 + \frac{(k-3)^2}{4})$$

распределена по Хи-квадрат с двумя степенями свободы.

Критерий Лиллиефорса — статистический тест, модификация критерия Колмогорова-Смирнова, для проверки гипотезы о согласии выборки с нормальным законом распределения. Особенностью являются изначально неизвестные параметры нормального распределения. Проводится в два этапа – сначала по выборочным оценкам строится нормальное распределение, затем применяется тест Колмогорова-Смирнова.

$D^* = \max_x \left|\hat{F}(x) - G(x)\right|$- статистика

$\hat{F}(x)$ –эмпирическая функция распределения

$G(x)$ – нормальное распределение, построенное по выборочному среднему и выборочной дисперсии

После производится сравнение критической и наблюдаемой величин и принимается. Либо отвергается гипотеза о согласии с нормальным распределением.

Критерий однородности Колмогорова – Смирнова – тест на принадлежность двух выборок одной генеральной совокупности.

Рассмотрим две выборки. Пусть эти выборки получены из генеральных совокупностей со следующими теоретическими распределениями - $F_1(x)$, $F_2(x)$. Естественно, теоретические функции распределения неизвестны, а в критерии применяются их эмпирические аналоги.

Выдвигается нулевая гипотеза $H_0: F_1(x) = F_2(x)$, в сравнении с альтернативной гипотезой $H_1: F_1(x) \neq F_2(x)$.

Статистика критерия Колмогорова-Смирнова подсчитывается следующим образом:

$$\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \max_x \left| F_{n_1}(x) - F_{n_2}(x) \right|$$

где $F_{n_1}(x)$, $F_{n_2}(x)$ – эмпирические функции распределения, определённые по соответствующим выборкам, величины выборок - n_1, n_2.

Статистика распределена по распределению Колмогорова.

Метод моментов — метод нахождения (оценки) неизвестного параметра распределения в математической статистике. Суть метода заключается в замене теоретического момента его выборочным аналогом.

Пусть $g(x) = (g_1(x), g_2(x) \dots g_l(x))$ - определённая функция, для которой задано математическое ожидание

$(\theta) = E\big(g(\xi)\big) = (Eg_1(\xi), Eg_2(\xi) \dots Eg_l(\xi))$.

А уравнение

$$m(\theta) = t \;\; t = (t_1, t_2 \dots t_l)\, m(\theta) = \big(m_1(\theta), m_2(\theta) \dots m_l(\theta)\big)$$

однозначно разрешимо относительно параметра θ, в области переменной $t \in m(\theta)$.

Пусть

$$\bar{g}\big(X_{[n]}\big) = \frac{1}{n}\sum_{i=1}^{n} g(x_i) \in m(\theta)$$

Тогда оценкой по методу моментов следует называть следующую величину:

$$\hat{\theta}\big(X_{[n]}\big) = m^{-1}(\bar{g}(X_{[n]})).$$

Нормальный закон распределения – закон распределения, задаваемый через функцию плотности вероятности Гаусса. Является одним из фундаментальных законов распределений в статистике.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \; e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\mu -$ параметр, а также математическое ожидание

σ — параметр распределения, кроме того среднеквадратичное отклонение

Математическое ожидание — мера среднего значения случайной величины. Подсчитывается, в дискретном и непрерывном случаях, следующим образом:

$$M[\zeta] = \begin{cases} \sum_{i} x_i p_i \\ \int_{-\infty}^{+\infty} f(x)dx \end{cases}$$

Дисперсия случайной величины— мера разброса случайной величины, её отклонение от математического ожидания.

$$D[x] = M[x_i - M[x]]$$

Медиана — значение, которое делит выборку на две равные части. Половина выборки будет меньше медианы, а половина – больше.

Мода — значение, наиболее часто встречающееся во множестве наблюдений. Иными словами, мода соответствует наблюдению с наибольшей частотой появлений.

Коэффициент эксцесса — число, характеризующее степень “остроты” графика распределения. Расчет коэффициента эксцесса выполняется по следующей формуле:

$$k = \frac{E(X - E[X])^4}{\sigma^4}$$

$E[X]$ — математическое ожидание

σ — стандартное отклонение

Для нормального распределения эксцесс равен трём, поэтому часто коэффициент эксцесса определяют, как $k - 3$, где k – вышеприведённый способ подсчёта. Чем больше коэффициент эксцесса, тем более вытянут вверх график распределения. Чем ближе к нулю, тем более ровный график.

Коэффициент асимметрии — число, характеризующее несимметричность графика распределения случайной величины, подсчитывается относительно центра - математического ожидания. Если коэффициент больше нуля – распределение смещено вправо, меньше нуля – влево. Симметричное распределение имеет нулевое значение коэффициента асимметрии. Расчёт коэффициента асимметрии производится следующим образом:

$$s = \frac{E(X - E[x])^3}{\sigma^3}$$

$E[X]$ — математическое ожидание

σ — стандартное отклонение

Линейная регрессионная модель — линейная модель, строящаяся по выборкам данных, для исследования взаимосвязи предикторов $x_1, x_2, x_3 \dots x_N$ и зависимой переменной y. Используется для исследований в регрессионном анализе. Коэффициенты линейной модели находятся с помощью метода наименьших квадратов.

$A \times X = B$ –уравнение для нахождения коэффициентов в матричном виде

$X = \begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_N \end{pmatrix}$ - неизвестные коэффициенты линейной модели

$$B = \begin{pmatrix} \sum_{i=1}^{M} y_i \\ \sum_{i=1}^{M} y_i x_{i,1} \\ \dots \\ \sum_{i=1}^{M} y_i x_{i,N} \end{pmatrix}$$

$$A = \begin{pmatrix} M & \sum_{i=1}^{M} x_{i,1} & \sum_{i=1}^{M} x_{i,2} & \sum_{\dots}^{\dots} \dots & \sum_{i=1}^{M} x_{i,N} \\ \sum_{i=1}^{M} x_{i,1} & \sum_{i=1}^{M} x_{i,1} x_{i,1} & \sum_{i=1}^{M} x_{i,2} x_{i,1} & \sum_{\dots}^{\dots} \dots & \sum_{i=1}^{M} x_{i,N} x_{i,1} \\ \sum_{i=1}^{M} x_{i,2} & \sum_{i=1}^{M} x_{i,1} x_{i,2} & \sum_{i=1}^{M} x_{i,} x_{i,} & \sum_{\dots}^{\dots} \dots & \sum_{i=1}^{M} x_{i,N} x_{i,2} \\ \sum_{\dots}^{\dots} \dots & \sum_{\dots}^{\dots} \dots & \sum_{\dots}^{\dots} \dots & \sum_{\dots}^{\dots} \dots & \sum_{\dots}^{\dots} \dots \\ \sum_{i=1}^{M} x_{i,N} & \sum_{i=1}^{M} x_{i,1} x_{i,N} & \sum_{i=1}^{M} x_{i,2} x_{i,N} & \sum_{\dots}^{\dots} \dots & \sum_{i=1}^{M} x_{i,N} x_{i,N} \end{pmatrix}$$

Коэффициент детерминации (R^2 , *R-квадрат, R-squared*) — это отношение, подсчитываемое для регрессионной модели. Интерпретируется, как доля дисперсии зависимой переменной, объясняемая рассматриваемой регрессионной моделью, то есть предикторами. Этот коэффициент используют как общепринятую меру зависимости одной случайной величины от множества других величин.

$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{RSS}{TSS}$ - “обычный” коэффициент детерминации

$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{(n-k)} \leq R^2$ – коэффициент детерминации, скорректированный на число коэффициентов. Данный коэффициент используется, когда необходимо сравнивать модели с различным числом признаков.

Проверка значимости линейной регрессии (F-test)

Проверка значимости линейной регрессии осуществляется посредством критерия Фишера. Статистика выглядит следующим образом:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F(k-1, n-k)$$

k - количество факторов в регрессионной модели, включая константу

n – объём рассматриваемой выборки

R^2 – коэффициент детерминации

В критерии проверяется гипотеза о равенстве нулю коэффициента детерминации. Если приведённая статистика больше критического значения, то регрессионная модель значима.

Коэффициент корреляции Пирсона – коэффициент корреляции, отражающий меру линейной связи двух выборок. Чем ближе коэффициент корреляции к единице, тем сильнее линейная взаимосвязь. Чем ближе коэффициент к нулю, тем больше данные говорят о линейной независимости.

Для следующих данных: $x^m = (x_1, x_2 \dots x_m), y^m = (y_1, y_2 \dots y_m)$ коэффициент корреляции вычисляется следующим образом:

$$r_{x,y} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})\ \sum_{i=1}^{m}(y_i - \bar{y})}}$$

- $\bar{x}, \bar{y}$– выборочные средние
- $r_{x,y} \in [-1,1]$.

Глава 3. Вероятностно-статистический анализ успеваемости

3.1 Вероятностно-статистический анализ зависимости успеваемости от факторов

Целью данного исследования является анализ многообразия представленных факторов, влияющих на успеваемость, в частности, прояснение значимости этих факторов численно. Рассматриваются следующие факторы модели: регион поступления, форма обучения (бюджет/вне бюджет), средний балл ЕГЭ, фактор зачисления студента по целевому набору, направление обучения, проходной балл ЕГЭ в год поступления, получение ста баллов по профильному предмету ЕГЭ по причине олимпиады определённого уровня, зачисление без экзаменов по причине олимпиады, особые обстоятельства поступления (сирота, инвалид и т.п.), форма обучения (дневная, вечерняя и т.п.).

До построения модели и корреляционно-регрессионного анализа проводится кодирование данных, представленных в текстовом виде. К примеру, результат зачисления студента без экзаменов приводится как название соответствующего профиля олимпиады: физика, математика или информатика, что следует переложить в численный вид. Само исследование проводится путём построения линейной регрессионной модели по имеющимся данным об успеваемости студентов и проверки гипотезы о равенстве нулю соответствующих коэффициентов.

Исследование проводится в системе MATLAB. Для построения линейной регрессионной модели используется встроенная функция fitlm.

Кроме построения модели данная функция проверяет гипотезу о значимости коэффициентов регрессии на основании t-критерия Стьюдента. Иными словами, для каждого из коэффициентов строится статистика и сравнивается с табличным значением. Таким образом, можно оценить степень влияния факторов на среднюю успеваемость.

Результаты построения модели следующие:

$$y = 3.85 + 0.0011x_1 - 0.4737x_2 + 0.0314x_3 - 0.1369x_4 - 0.0001x_5 - 0.004\,x_6 - 0.0385\,x_7 + 0.1049x_8 + 0.1070x_9 + 0.3557x_{10}$$

Коэффициент детерминации модели - 16.2%, при величине выборки 31020 значения. Согласно критерию Фишера получаем значимость коэффициента детерминации с 99% уровнем, и, как следствие, значимость регрессионной модели, F=621,45 > $F_{\text{критического}}$=2,32.

Фактор	**Уровень значимости**	**Переменная**
Регион поступления	96.9%	x1
Форма обучения (бюджет/коммерция)	99.93%	x2
Средний балл ЕГЭ	~100%, (1-8,17e-63)*100%	x3
Целевой набор	61.39%	x4
Направление обучения	95.37%	x5
Проходной балл ЕГЭ	~100.00% (1-1,01e-10)*100%	x6

Олимпиада (100 баллов по ЕГЭ)	97.85%	x7
Олимпиада (зачисление без экзаменов)	99.94%	x8
Особые обстоятельства зачисления	99.49%	x9
Форма обучения (дневная, вечерняя, заочная)	~100% (1-1,30e-11)*100%	x10

В качестве вывода, факторы, исходя из таблицы, группируются следующим образом:

1. Незначимый фактор, имеется в виду значимость менее 95%,- целевой набор
2. Условно-значимые, с уровнем значимости 95%-99.9%,- регион поступления, направление обучения, фактор олимпиады (100 баллов по ЕГЭ), особые обстоятельства зачисления
3. Значимые, с уровнем значимости чуть выше 99.9%,- форма обучения (бюджет/вне бюджета), олимпиада (зачисление без экзаменов)
4. Высокая значимость, уровень значимости порядка $(100\text{-}10^{-9})\%$,- проходной балл ЕГЭ, форма обучения (дневная, вечерняя или заочная)
5. Особо значимые, уровень порядка $(100\text{-}10^{-61})\%$,- средний балл ЕГЭ

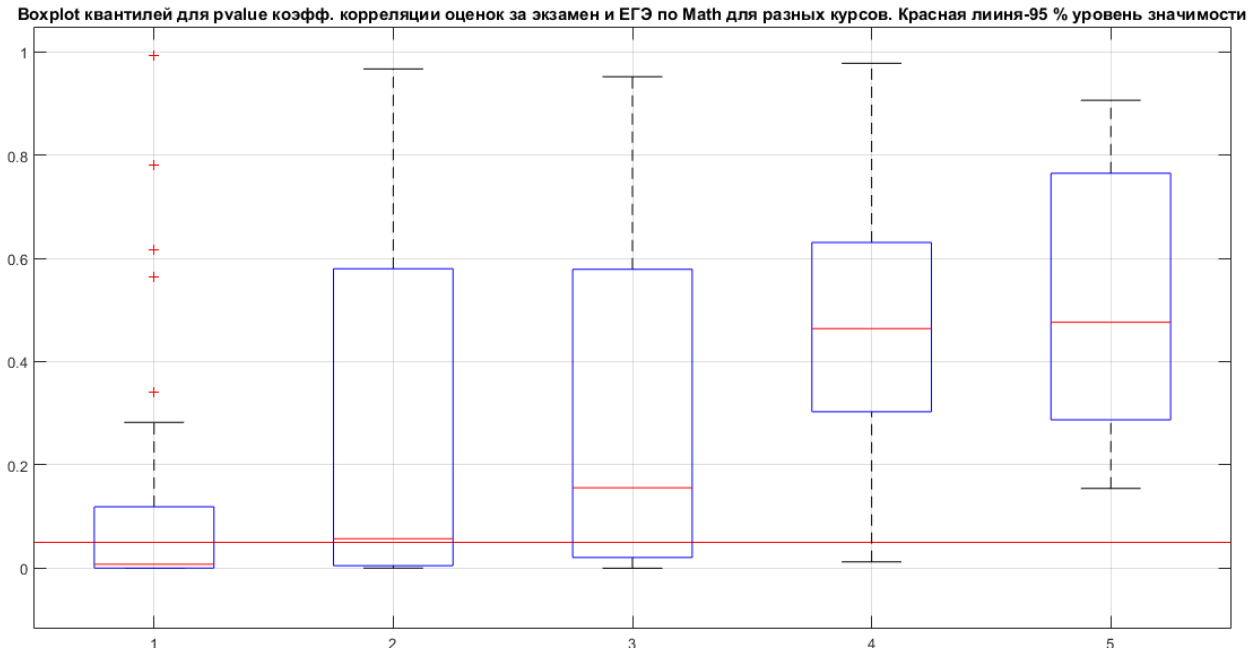
3.2 ЕГЭ и успеваемость

Данный параграф призван оценить прогностическую валидность баллов ЕГЭ. Исследование проводится путём построения многофакторной регрессионной модели и рассмотрением коэффициента детерминации этой

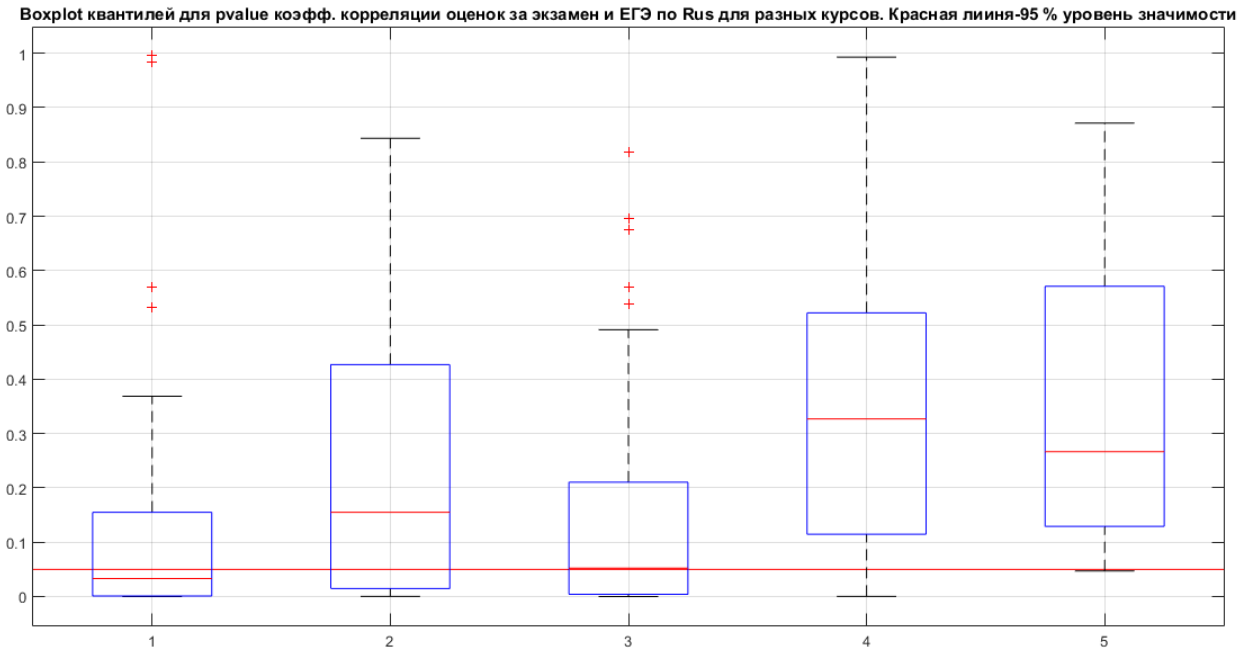
модели. Похожие исследования были проведены Т. Е. Хавенсоном и А. А. Соловьевой в работе “Связь результатов Единого государственного экзамена и успеваемости в вузе” [5]. В частности, статья говорит о том, что достаточная прогностическая валидность баллов ЕГЭ находится в интервале от 0,35 до 0,46 . Иначе говоря, отметки гос. экзамена предсказывают 12–25% вариации успеваемости в вузе (величина R^2).

На нижеследующих графиках рассмотрена корреляция баллов ЕГЭ и оценок за экзамен с сечением по курсам. Это делается с целью выявить необходимые для включения в регрессионную модель факторы и оценить необходимость рассмотрения валидности баллов ЕГЭ для разных курсов обучения. Для построения графиков использовался коэффициент корреляции Пирсона и его статистика. На графике отображено p-value статистики. Для построения использовались функции пакета MATLAB corr и boxplot. Для вычисления p-value использовалась функция corr. Для построения графика boxplot.

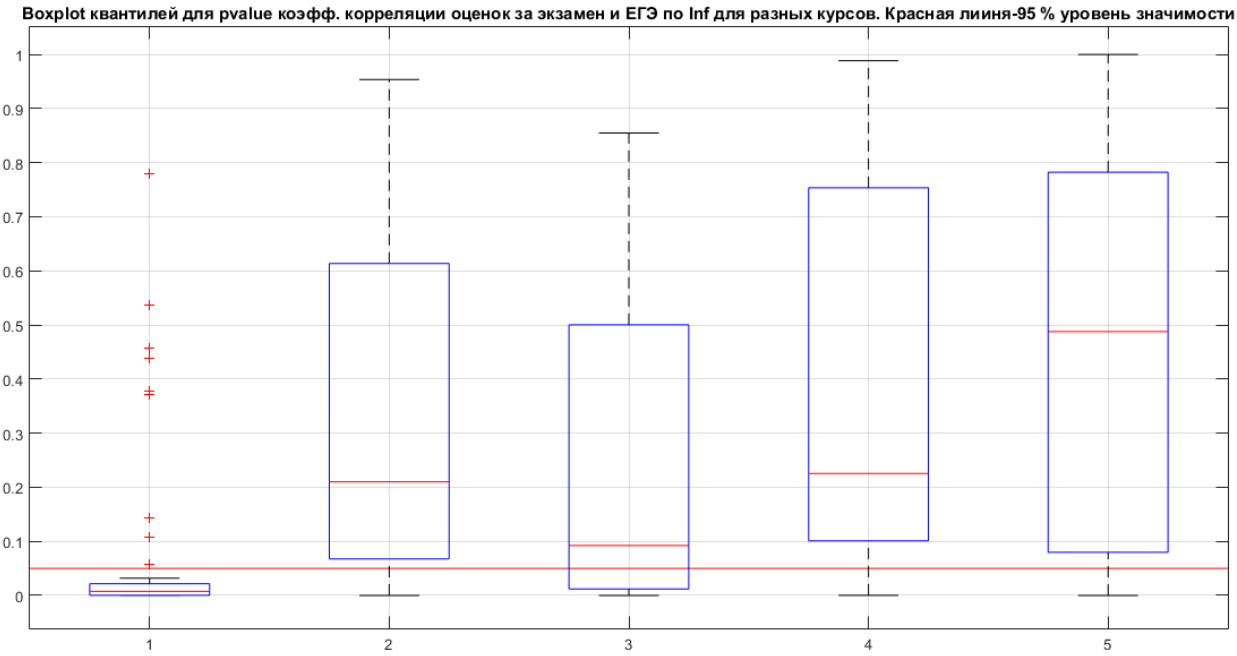
Для всех графиков красная линия означает 5% порог. Если значение ниже, то для него отвергается гипотеза о нулевом значении коэффициента корреляции. Иными словами, баллы ЕГЭ и отметки успеваемости по предмету коррелируют.



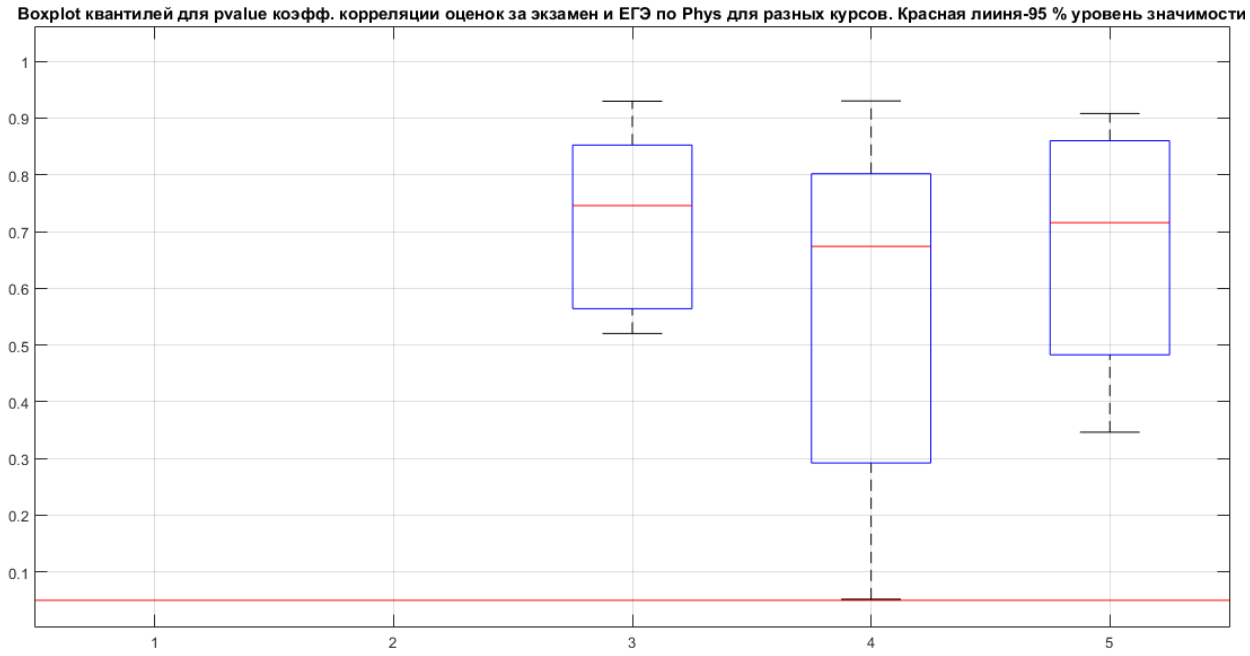
Из приведённого графика следует вывод, что более чем в половине случаев для первокурсников взаимосвязь баллов ЕГЭ и оценок значима. Для студентов второго курса это верно чуть менее чем в половине случаев. Порядка четверти случаев выставления отметок по экзамену на третьем курсе можно объяснить баллами ЕГЭ. Для четвёртого и пятого курсов число объяснений баллами гос. экзамена выставления отметок становится несущественным.



Из данного графика можно заключить, что существенными к рассмотрению в общей модели линейной регрессии являются первый, второй и третий курсы, в случае объяснения успеваемости баллами по русскому языку.



Из вышеприведённого графика следует вывод, что в подавляющем большинстве случаев для первокурсников взаимосвязь баллов ЕГЭ и оценок значима. Но эта значимость не сохраняется более чем в 75% эпизодах выставления отметок уже на втором курсе. На третьем курсе случаев объяснения баллами по информатике выставления отметок становится немногим больше. Количество же таких случаев на четвёртом и пятом курсе незначительно.



В силу вышеприведённого графика, баллы ЕГЭ по физике не включаются в рассмотрение для построения регрессионной модели и оценки прогностической валидности баллов ЕГЭ. Эта картина корреляционной зависимости баллов ЕГЭ по физике и успеваемости объясняется набором исходных данных и условием конкурса на поступление в разное время.

Рассуждения, приведённые выше, говорят о том, что в общую регрессионную модель стоит включить математику, русский и информатику на первом курсе. На втором курсе отметки следует объяснять баллами по

математике и русскому. На третьем - математикой, русским и информатикой. Взаимосвязь баллов ЕГЭ и успеваемости на четвёртом и пятом курсах рассматривать нет необходимости. Общая динамика объяснимости говорит о том, что в модель следует также включить баллы ЕГЭ по информатике на втором курсе. Поэтому ниже будет рассмотрен противовес двух регрессионных моделей.

Сложно оценить корреляцию проходного балла и среднего балла студента за экзамен в силу того, что проходной балл для студентов одной укрупнённой группы в определённую сессию одинаков. Исключения составляют разве что переведённые и восстановленные студенты, которые составляют незначительную долю студентов. Однако проходной балл может быть информативен для построения общей регрессионной модели. Исходя из этого, регрессионные модели будут строиться как с учётом проходного балла ЕГЭ, так и без него, а также с проверкой гипотезы о равенстве нулю коэффициента модели при проходном балле.

Регрессионная модель строится специальной функцией MATLAB fitlm. Для корректного сравнения моделей будем использовать скорректированный коэффициент детерминации. Он всегда меньше обычного коэффициента детерминации. Он трактуется как доля объяснимой регрессии не меньше чем R^2-adjusted. Ниже построены линейные модели (слева) и коэффициент детерминации (справа). Также приведены значения pValue. Если значение меньше 0.05, то с 95% значимостью следует отвергнуть гипотезу о равенстве нулю коэффициента при соответствующей переменной.

В регрессионной модели будут использоваться следующие переменные:

- y- средний балл студента за сессию
- m- баллы ЕГЭ по математике
- i- баллы ЕГЭ по информатике
- r - баллы ЕГЭ по русскому
- p- проходной балл в год поступления

Модель для ЕГЭ и успеваемости 1 курса

$$y = 1.6997 + 0.012\,m + 0.0141\,i + 0.0125\,r - 0.0043\,p$$

	Свободный член	Математика	Информатика	Русский	Проходной балл	R^{2*}	Выборка
Коэффициенты модели	1.6997	0.012	0.0141	0.0125	-0.0043	**20%**	6451
pValue	$3.06*10^{-103}$	$7.12*10^{-87}$	$1.07*10^{-61}$	$1.5*10^{-72}$	$1.76*10^{-39}$		

*Регрессия значима с 99% уровнем, по критерию Фишера - $F= 402.875 > F_{критического}=3.32$

Модель для ЕГЭ и успеваемости 1 курса без учёта проходного балла

$$y = 1,3394 + 0,0127m + 0,0098i + 0,0094r$$

	Свободный член	Математика	Информатика	Русский	R^{2*}	Выборка
Коэффициенты модели	1,3394	0,0127	0,0098	0,0094	**18%**	6451
pValue	$1,45*10^{-72}$	$8,17*10^{-95}$	$2,55*10^{-35}$	$1,41*10^{-46}$		

*Регрессия значима с 99% уровнем, по критерию Фишера - $F=471.73 > F_{критического}=3.78$

Вышеприведённые модели говорят нам о том, что баллы ЕГЭ для студентов первого курса имеют достаточную прогностическую валидность в объяснении успеваемости, об этом говорит то, что коэффициенты

детерминации более 12%. Также следует отметить, что гипотезы о равенстве нулю коэффициентов при соответствующих отметках баллов ЕГЭ мы отвергаем с 99% уровнем значимости. Иначе говоря, все результаты гос. экзамена значимы в объяснении академического прогресса студента. Что и подтверждается различием коэффициента детерминации первой и второй модели.

Резюмируя вышесказанное, баллы ЕГЭ по математике, информатике и русскому, а также проходной балл в год поступления объясняют 20% вариации оценок в ВУЗе на первом курсе.

Модель для ЕГЭ и успеваемости 2 курса

$$\boldsymbol{y = 1.5284 + 0.0084m + 0.0111i + 0.0148r - 0.0011p}$$

	Свободный член	Математика	Информатика	Русский	Проходной балл	R^{2*}	Выборка
Коэффициенты модели	1.5284	0.0084	0.0111	0.0148	-0.0011	**16%**	5716
pValue	$2.52*10^{-51}$	$1.20*10^{-33}$	$3.01*10^{-27}$	$2.3*10^{-74}$	0.0032		

*Регрессия значима с 99% уровнем, по критерию Фишера - $F=271.95 > F_{критического}=3,32$

Модель для ЕГЭ и успеваемости 2 курса без информатики

$$\boldsymbol{y = 1.9865 + 0.0107m + 0.0152r + 0.00016p}$$

	Свободный член	Математика	Русский	Проходной балл	R^{2*}	Выборка
Коэффициенты модели	1.9865	0.0107	0.0152	0.00016	**13%**	5716

pValue	$1.13*10^{-98}$	$1.8*10^{-57}$	$2.87*10^{-76}$	**0.64**

*Регрессия значима с 99% уровнем, по критерию Фишера - $F=284,5 > F_{критического}=3,78$

Модель для ЕГЭ и успеваемости 2 курса без проходного балла

$$y = 1.4225 + 0.0086m + 0.0101i + 0.0142r$$

	Свободный член	Математика	Информатика	Русский	R^{2*}	Выборка
Коэффициенты модели	1.4225	0.0086	0.0101	0.0142	**15%**	5716
pValue	$6.3*10^{-31}$	$3.95*10^{-35}$	$2.19*10^{-25}$	$1.71*10^{-73}$		

*Регрессия значима с 99% уровнем, по критерию Фишера - $F=336 > F_{критического}=3,78$

Модель для ЕГЭ и успеваемости 2 курса без проходного балла и информатики

$$y = 2.0103 + 0.0107m + 0.0153r$$

	Свободный член	Математика	Русский	R^{2*}	Выборка
Коэффициенты модели	2.0103	0.0107	0.0153	**13%**	5716
pValue	$1.5*10^{-144}$	$1.63*10^{-57}$	$1.71*10^{-84}$		

*Регрессия значима с 99% уровнем, по критерию Фишера - $F=426,83 > F_{критического}=4,6$

Линейные регрессионные модели для результатов экзамена при поступлении и отметок успеваемости в ВУЗе на втором году обучения также проходят через 12% порог, что демонстрирует их значимость. Коэффициенты детерминации этих моделей существенно ниже, чем на первом курсе. Лучшая из приведённых моделей демонстрирует нам 16%-ю долю

объяснимой вариации против 20% для модели на первом курсе. Предположения о равенстве нулю коэффициентов при соответствующих отметках баллов ЕГЭ мы отвергаем с 99% уровнем значимости. Иными словами, все результаты гос. экзамена значимы в объяснении академического прогресса студента. Следует отметить, что исключение из рассмотрения проходного балла привело к ухудшению модели на 1% при значимости коэффициента (1-0.0032)*100%. Такого же рода исключение привело к ухудшению результата на 2% в первом случае, но значимость коэффициента была выше – $(1-10^{-39})*100\%$. Также необходимо отметить модель для ЕГЭ и успеваемости 2 курса без информатики в сравнении с моделью без учёта, как информатики, так и проходного балла. Различие этих моделей заключается в учёте проходного балла для прогностической валидности. В первой модели гипотезу о нулевом значении коэффициента при проходном балле мы принимаем и, как демонстрирует нам нулевое различие их коэффициентов детерминации, это упрощает модель без изменения её качества.

Различие в значимости коэффициентов регрессионной модели приводят к схожим тенденциям в различии коэффициента детерминации.

Исключение фактора, не прошедшего проверку гипотезой о неравенстве нулю коэффициента в регрессионной модели, действительно не ухудшает качества модели.

Учёт оценки единого экзамена по информатике на втором курсе для построения регрессии, как и предполагалось, оказался информативен.

В качестве заключения, отметки государственного оценивания по математике, информатике и русскому в совокупности с проходным баллом на втором курсе объясняют 16% вариации успеваемости.

Модель для ЕГЭ и успеваемости 3 курса

$$y = 1.5318 + 0.0045m + 0.0119i + 0.0159r + 0.00013p$$

	Свободный член	Математика	Информатика	Русский	Проходной балл	R^{2*}	Выборка
Коэффициенты модели	1.5318	0.0045	0.0119	0.0159	0.00013	**15%**	5241
pValue	$3.04*10^{-39}$	$2.25*10^{-8}$	$3.60*10^{-25}$	$2.3*10^{-48}$	**0.77**		

*Регрессия значима с 99% уровнем, по критерию Фишера - $F=231 > F_{критического}=3,32$

Модель для ЕГЭ и успеваемости 3 курса без проходного балла

$$y = 1.5469 + 0.0045m + 0.0120i + 0.0159r$$

	Свободный член	Математика	Информатика	Русский	R^{2*}	Выборка
Коэффициенты модели	1.5469	0.0045	0.0120	0.0159	**15%**	5241
pValue	$4.53*10^{-49}$	$2.32*10^{-08}$	$4.11*10^{-27}$	$1.9*10^{-50}$		

*Регрессия значима с 99% уровнем, по критерию Фишера - $F=308,05 > F_{критического}=3,78$

Моделирование успеваемости на 3 курсе результатами единых вступительных испытаний имеет прогностическую возможность в 15%. Это меньше чем в предыдущих случаях, из чего можно сделать заключение о

пропорциональности коэффициента детерминации и курса обучения. Проходной балл для построения регрессионной модели незначим. Первая модель демонстрирует это превышением p-значения порога в 0.05 единиц, а вторая модель подтверждает коэффициентом детерминации. Остальные коэффициенты регрессии значимы.

Исключение из рассмотрения проходного балла не ухудшает модель.

Отметки баллов ЕГЭ по математике, информатике и русскому на третьем курсе объясняют 15% вариации успеваемости.

Общие выводы:

- Модели, которые успешно предсказывают успеваемость, и соответствующие коэффициенты детерминации имеют вид:

1 курс: $y = 1.6997 + 0.012\,m + 0.0141\,i + 0.0125\,r - 0.0043\,p$ **20%**

2 курс: $y = 1.5284 + 0.0084m + 0.0111i + 0.0148r - 0.0011p$ **16%**

3 курс: $y = 1.5469 + 0.0045m + 0.0120i + 0.0159r$ **15%**

- Соглашаясь со статьёй [5], данные говорят о том, что отметками единого вступительного экзамена можно объяснить вариацию успеваемости на первом курсе. Причём, доля объяснимой вариации составляет 20%, что также совпадает с результатом работы Хавенсона и Соловьёвой.
- Модели, построенные для количественного учёта взаимосвязи баллов ЕГЭ и академического прогресса для второго и третьего курса, не соглашаются с приведённой работой. Эти модели показывают значимые коэффициенты детерминации и коэффициенты регрессионной модели

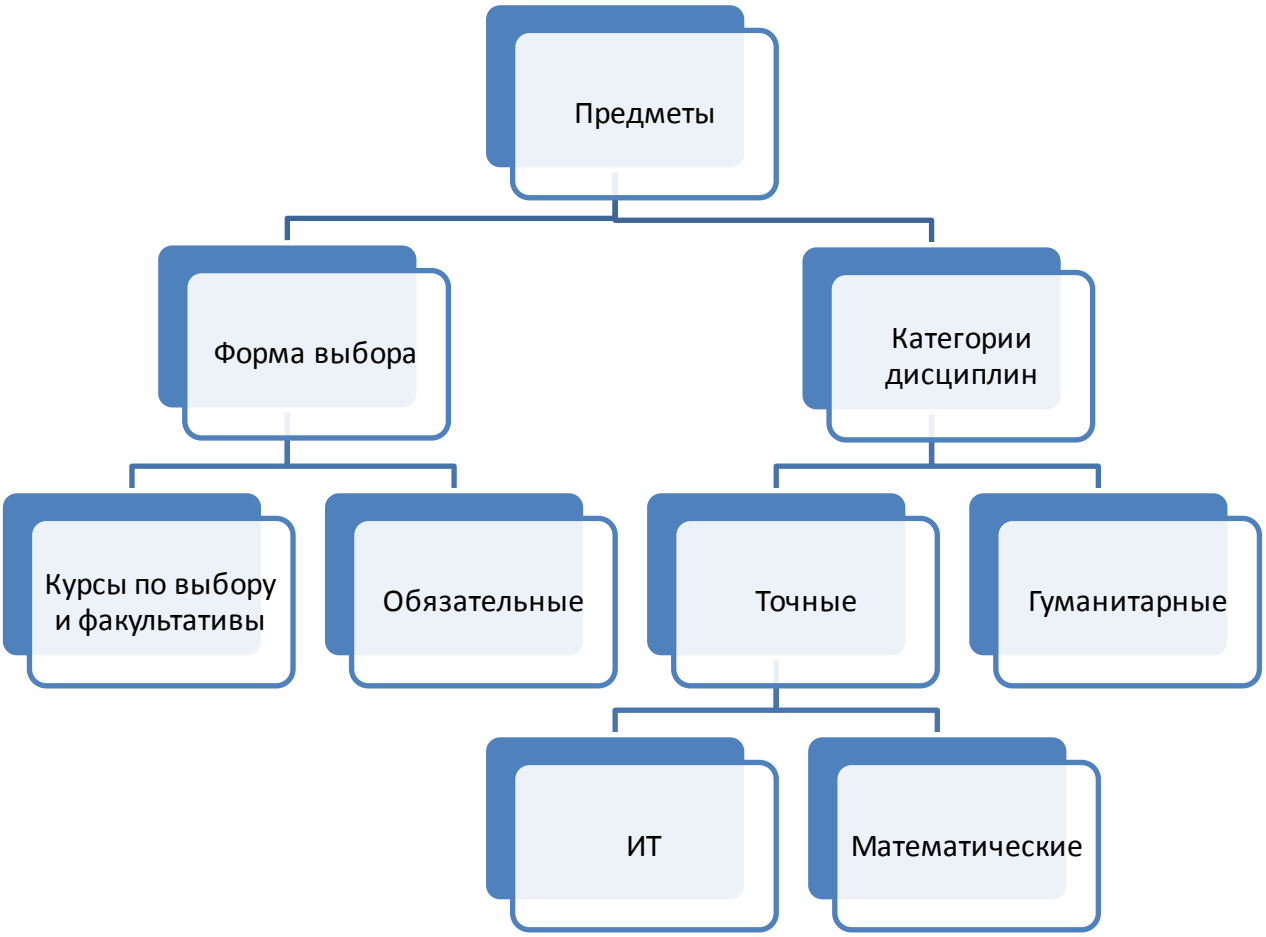
- Объяснение успеваемости отметками ЕГЭ с курсом обучения понижается

- Различие в значимости коэффициентов регрессионной модели приводят к схожим тенденциям в различии коэффициента детерминации

- Значимость проходного балла с курсом понижается, пока на третьем курсе не появляется возможность его исключения из рассмотрения

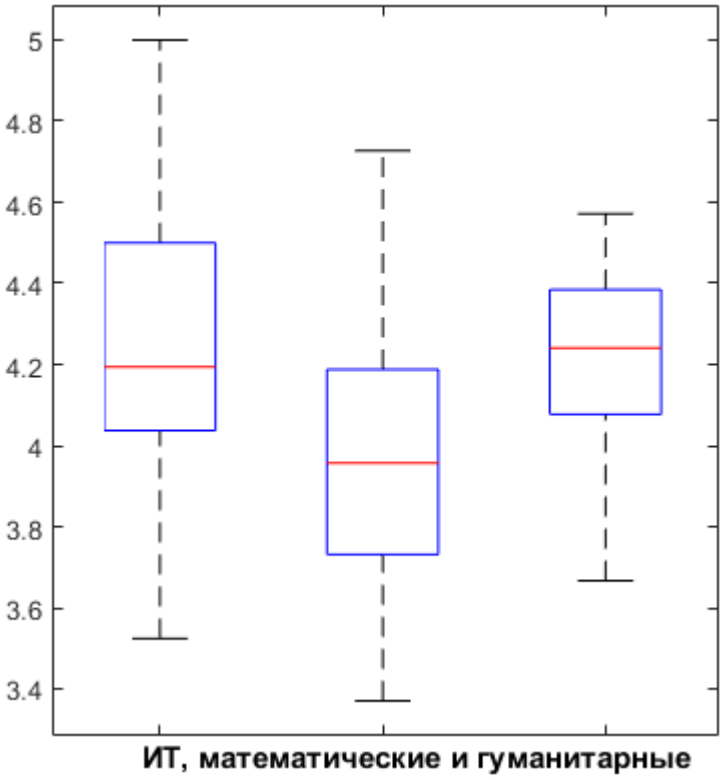
3.3 Предметы и успеваемость

Предметы, изучаемые в ВУЗе, бывают разные. Гуманитарные и точные, курсы по выбору и факультативы. Нижеприведённые выкладки проясняют количественными статистическими исследованиями качественные связи между группами дисциплин. Схема, изображённая далее, отражает группировку предметов по различным категориям и взаимосвязь этих групп. Исследование построено на сравнении успеваемости по соответствующим категориям предметов с помощью статистического аппарата и прикладного пакета программ.



Ниже приведён “Ящик с усами” для сравнения успеваемости по разным категориям дисциплин: информационным технологиям, дисциплинам, в большей степени относящимся к чистой математике, и гуманитарным наукам. Например, к информационным технологиям относятся “Основы программирования”, “Языки программирования” и “Архитектура вычислительных систем”. Некоторые исключительно математические предметы - “Математический анализ”, “Алгебра” и “Дифференциальные уравнения”. К гуманитарным предметам были отнесены такие дисциплины, как “Философия”, “Иностранный язык” и “История”. Диаграмма построена с помощью системы MATLAB и встроенной

функции Boxplot, вычисляющей 25% 50% и 75% перцентили. Нижняя и верхняя границы строятся в соответствии с правилом трёх сигм.



Как мы видим из построенного выше графика, успеваемость по гуманитарным и ИТ предметам в целом выше, чем по математическим. Это может быть охарактеризовано сложностью изучения данных дисциплин студентами. Также данные говорят о том, что разброс оценок по гуманитарным предметам меньше, чем по математическим и ИТ предметам. Что можно описать, как боле усреднённый уровень успешности в освоении гуманитарных дисциплин. Иначе говоря, освоение гуманитарных дисциплин характеризуется большей однородностью.

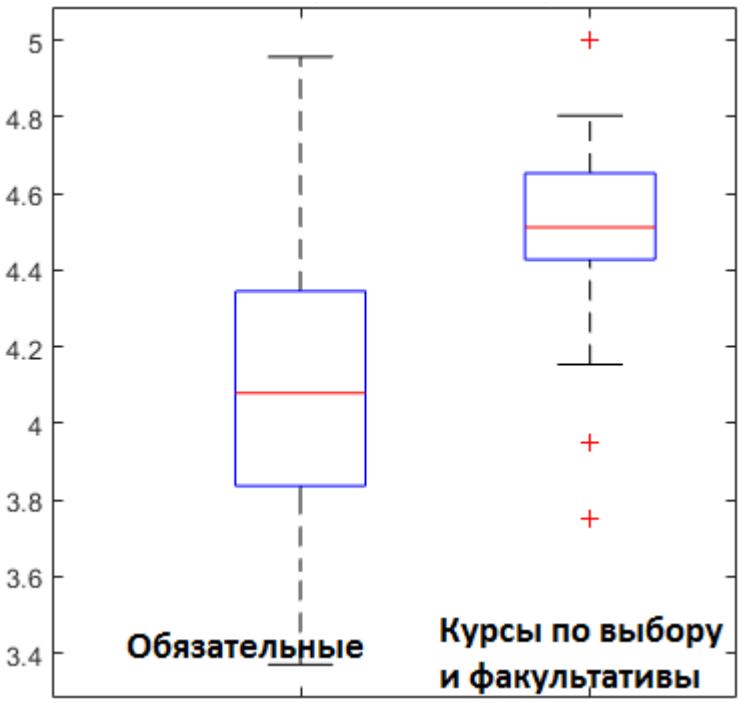
Разнородность оценок по дисциплинам говорит о необходимости выдвижения гипотезы об однородности данных. Воспользуемся тестом Колмогорова-Смирнова для проверки гипотезы. В пакете программ MATLAB

данный тест реализован в виде функции kstest2. Также воспользуемся сводной описательной таблицей DescrStat, в которой соответствующие категории дисциплин закодированы определённым образом.

С 99.9% уровнем значимости гипотеза, что уровни успеваемости по математическим дисциплинам и предметам, изучающим информационные технологии, распределены по одному генеральному закону, не отвергается. Иными словами, успеваемости по информационным технологиям и математическим предметам статистически однородны. В отличие от успеваемости по математическим и информационно-технологическим дисциплинам в сравнении с гуманитарными предметами.

Ниже следует сравнение обязательных к выбору дисциплин и необязательных. К первым относятся такие дисциплины, как “Численные методы”, “Теория вероятности и математическая статистика” и “Геометрия”. Ко вторым категориям дисциплин относятся курсы по выбору и факультативы. Например, “Курс по выбору: пакет Simulink, сетевые технологии или математическое моделирование” или необязательный к изучению факультатив “Методы управления в социально-экономических системах”.

Далее приведён график для сравнения средних уровней успеваемости по соответствующим категориям дисциплин. Схема создана с помощью пакета прикладных программ MATLAB и встроенной в неё функции boxplot, вычисляющей 25% 50% и 75% перцентили. Нижняя и верхняя границы графика строятся в соответствии с правилом трёх сигм.



Вышеприведённая схема говорит о том, что уровень успеваемости по необязательным к выбору предметам в целом выше, чем по остальным. Значимо ли отличие статистически, взяты ли выборки из одной генеральной совокупности - скажет тест Колмогорова-Смирнова и его реализация в системе MATLAB kstest2. Также, для исследования будем использовать сводную описательную таблицу DescrStat, в которой соответствующие предметы закодированы и разграничены.

С 95% уровнем значимости, гипотеза об однородности успеваемости по дисциплинам, обязательным к изучению и необязательным, не отвергается. Иными словами, факторы, влияющие на успеваемость по обязательным дисциплинам, статистически равно влияют и на успеваемость по дисциплинам, не обязательным к выбору.

Вывод:

1. Успеваемости по предметам информационных технологий и математическим дисциплинам однородны

2. Успеваемость по математическим предметам (или дисциплинам информационных технологий) и гуманитарным неоднородна

3. Курсы, обязательные и необязательные к изучению, характеризуются единым набором факторов, в равной степени, влияющей на их успеваемость

3.4 Проверка гипотезы о распределении успеваемости

Нормальное распределение играет важную роль в статистике, обработке и анализе данных. При условии согласия выборки с нормальным распределением существует множество эффективных методик для описания и анализа данных. К данной проблематике относятся работы [2, 12] . В работе [2] был получен результат о согласии оценок по математическому анализу с нормальным распределением $[x_i]$ ~N (3.5; 1, 0). Работа [12] также подтверждает согласие с нормальным распределением успеваемости, в частности тестирования по химии.

Ниже рассматривается средний балл студента в сессию, взятый по всем экзаменам**.** Гистограмма строится средствами графики пакета прикладных программ MATLAB функцией histfit. Эта функция строит гистограмму распределения успеваемости, а красной линией изображает нормальное распределение с параметрами, оценёнными по выборке с помощью метода моментов.

Выдвигается гипотеза о согласии среднего балла с нормальным распределением. Функция jbtest системы MATLAB подсчитает результат о согласии данных с нормальным распределением, при неизвестных параметрах, с помощью теста Ярки-Бера. С 99.9% значимостью мы не отвергаем гипотезу о нормальности распределения среднего балла студента в сессию.

$$P\left([x_i] \sim N(\mu; \sigma^2)\right) = 99.9\%$$

После согласия выборки с распределением необходимо узнать параметры этого распределения. Для нахождения значений параметров служит функция из пакета Statistics toolbox программы MATLAB normfit. В ней

используется метод моментов для нахождения неизвестных значений распределения. Кроме точного нахождения неизвестных, также строятся доверительные интервалы для этих параметров с 95% уровнем значимости. Для их построения используются соответствующие статистики. Ниже приведены результаты подсчёта:

$$[x_i] \sim N(4.0597; 0.6170^2)$$

$$P(\mu \in \langle 4.038 | 4.0811 \rangle) = 95\%$$

$$P(\sigma \in \langle 0.6023 | 0.6325 \rangle) = 95\%$$

Вышеприведённое исследование ставит под вопрос распределение средних оценок студента в сессию. Ниже рассматривается аналогичный вопрос о согласии с нормальным распределением, но уже оценок по всем экзаменам $[y_i]$. Исследование, по аналогии с вышеизложенным, воспользуется тестом Ярки-Бера, методом моментов, соответствующими статистиками для построения доверительных интервалов и функциями MATLAB: jbtest,normfit.

С 99.9% уровнем значимости принимается гипотеза о не противоречии отметок за экзамен нормальному распределению. Ниже следуют параметры распределения и соответствующие доверительные интервалы:

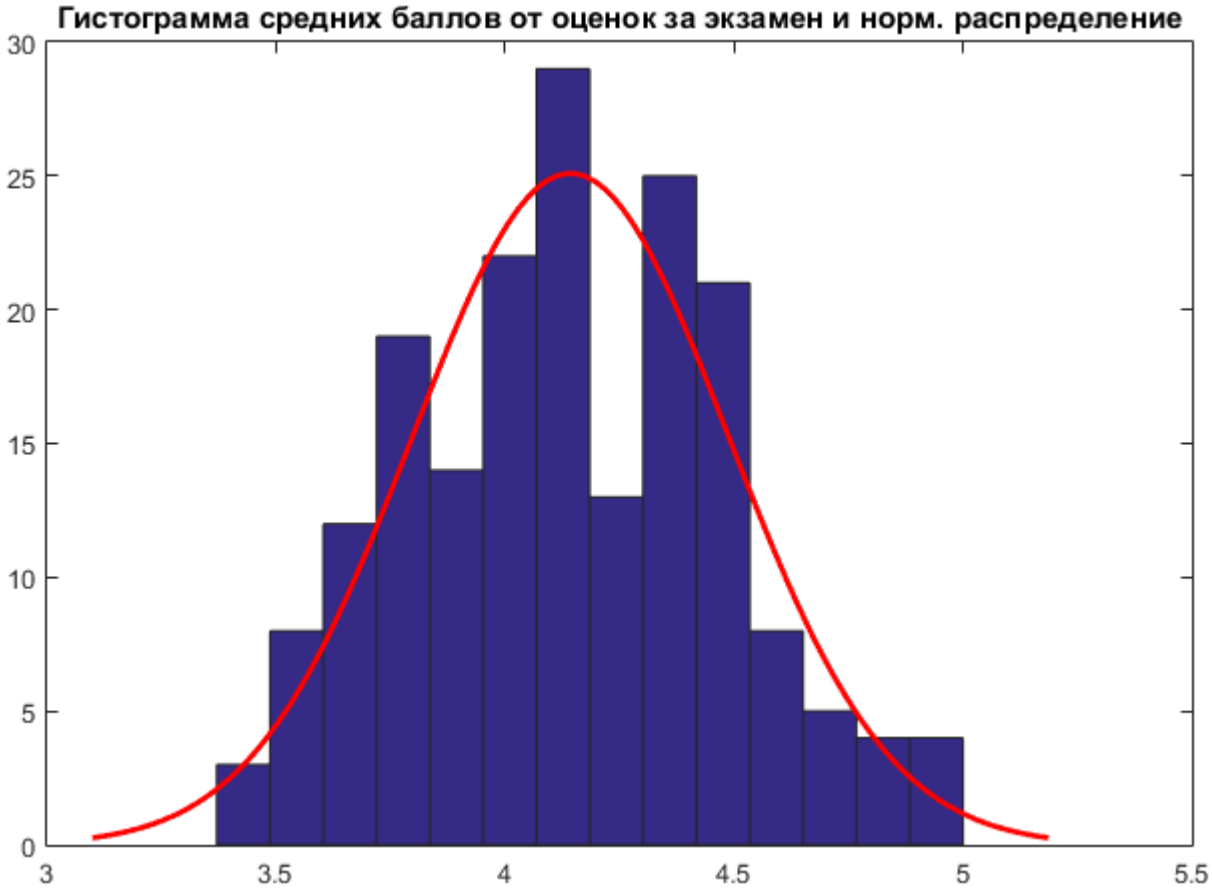
$$P\left([y_i] \sim N(\mu; \sigma^2)\right) = 99.9\%$$

$$[y_i] \sim N(4.0537; 0.8247^2)$$

$$P(\mu \in \langle 4.0399 | 4.0675 \rangle) = 95\%$$

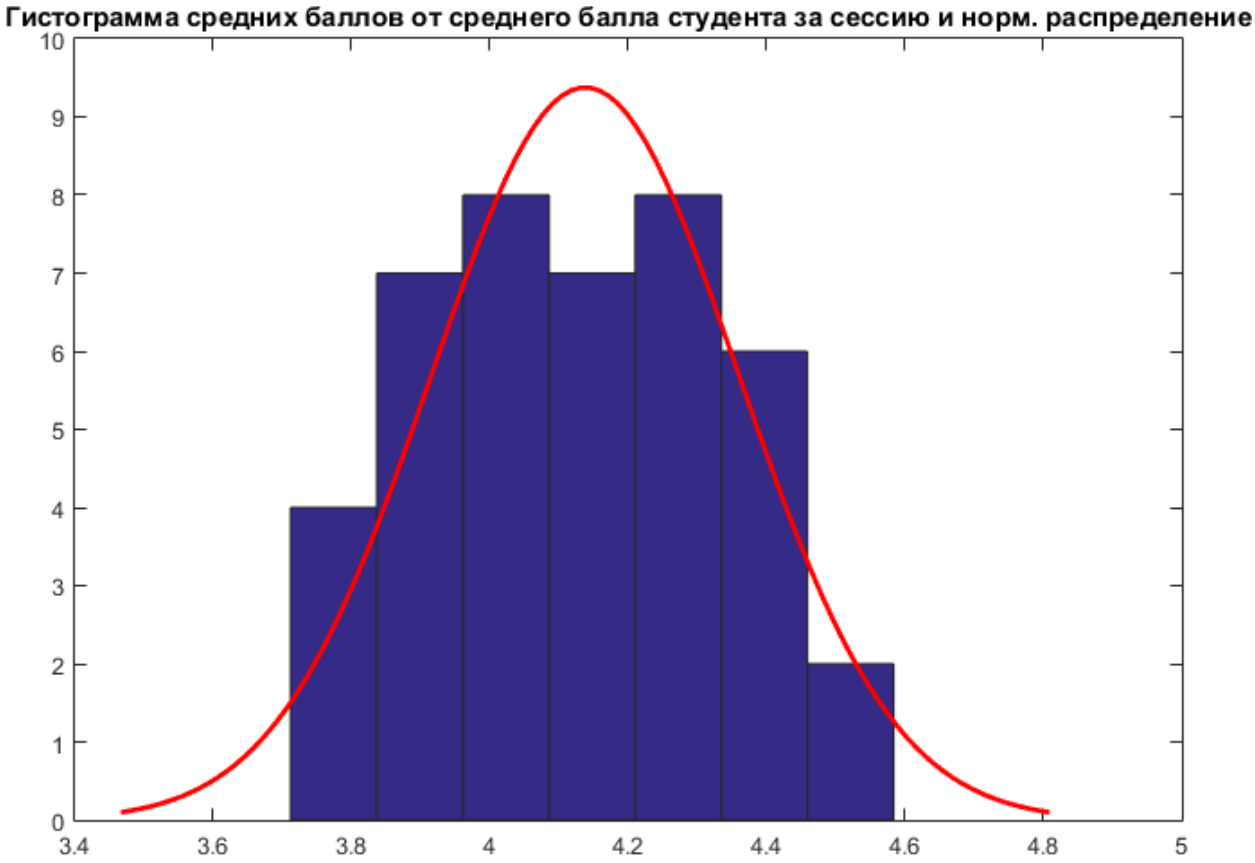
$$P(\sigma \in \langle 0.8150 | 0.8345 \rangle) = 95\%$$

Далее проводится исследование значений средних баллов от величин, исследуемых выше. Средний балл от оценок за экзамен и средняя величина от среднего балла студента в сессию.



Распределение средних баллов за экзамен с 95% уровнем значимости по тесту Ярки-Бера, точнее его реализации в системе MATLAB jbtest, отвергает согласованность с нормальным распределением. Уровень значимости, при котором гипотеза не была бы отвергнута, составляет 62.5% .

$$P\left(\bar{x}\sim N(\mu;\sigma^2)\right) = 62.5\%$$



Распределение среднего балла от средних баллов студента за сессию не согласуется с нормальным распределением при уровне значимости 95%.Уровень значимости, при котором гипотеза не будет отвергнута, составляет 60.3% .

$$P\left(\bar{y}\sim N(\mu;\sigma^2)\right) = 62.5\%$$

Результаты вышеприведённых исследований разнятся. Возможная причина кроется в малой размерности выборок при исследовании средних величин. Это 182 значения для $\bar{x}$, среднего балла за экзамен и 42 для $\bar{y}$, среднего балла от средних баллов студентов в сессию. К сравнению, в исследовании о согласии среднего баллаот средних баллов с нормальным распределением $[x_i]\sim N$ выборка исчислялась тысячами значений, а в

согласии оценок за экзаменс нормальным распределением $[y_j] \sim N$ – десятками тысяч значений.

Выводы:

1. Для исследования распределения средних баллов от отметок за экзамен и среднего балла студента в сессию необходимо увеличить размер выборки

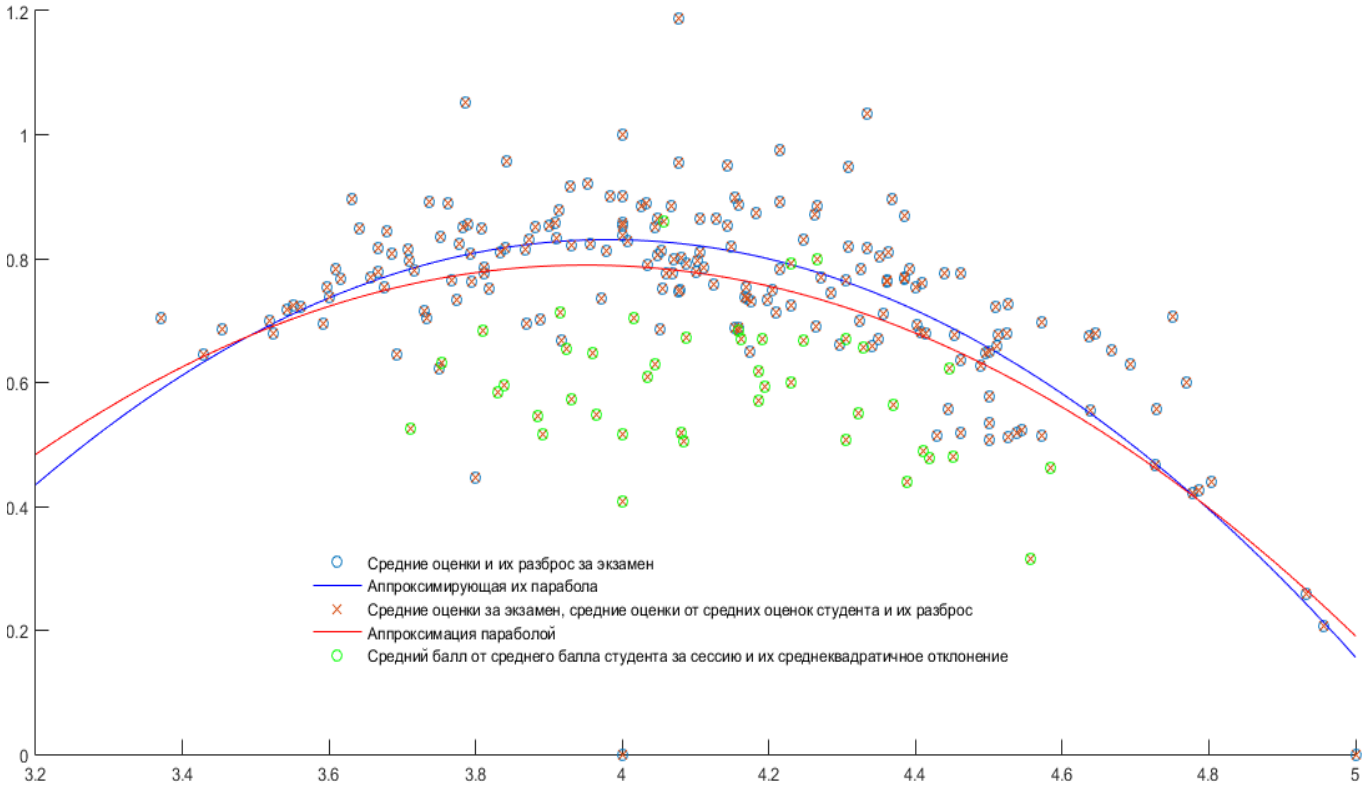
2. Отметки за экзамен согласуются с нормальным распределением
$$[y_i] \sim N(4.0537; 0.8247^2)$$

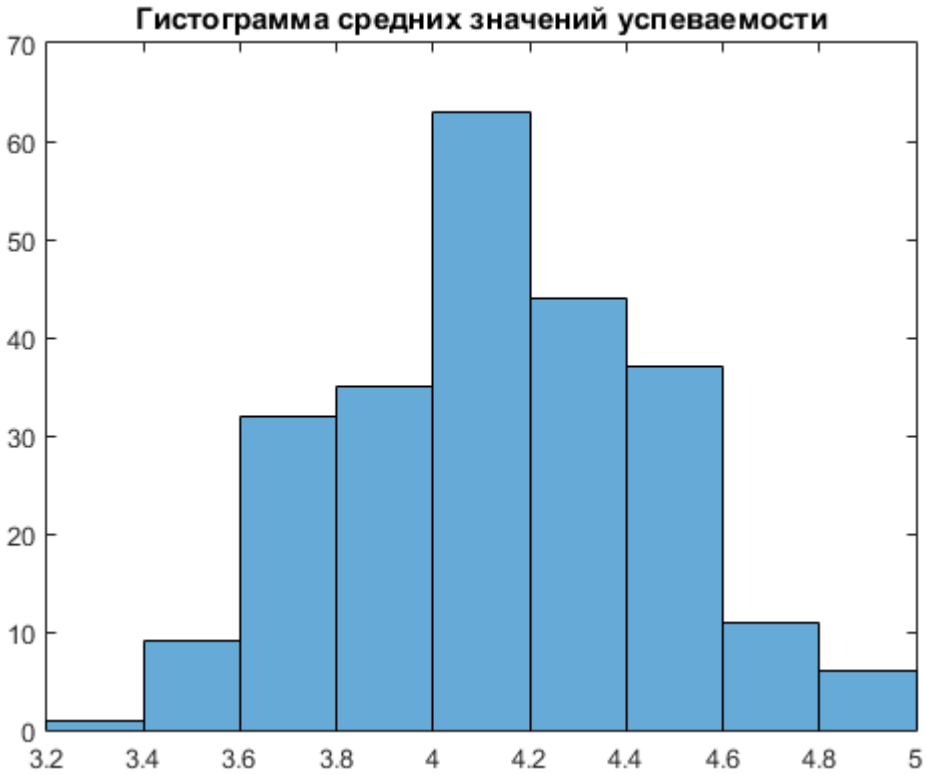
3. Средний балл студента за сессию согласуется с нормальным распределением
$$[x_i] \sim N(4.0597; 0.6170^2)$$

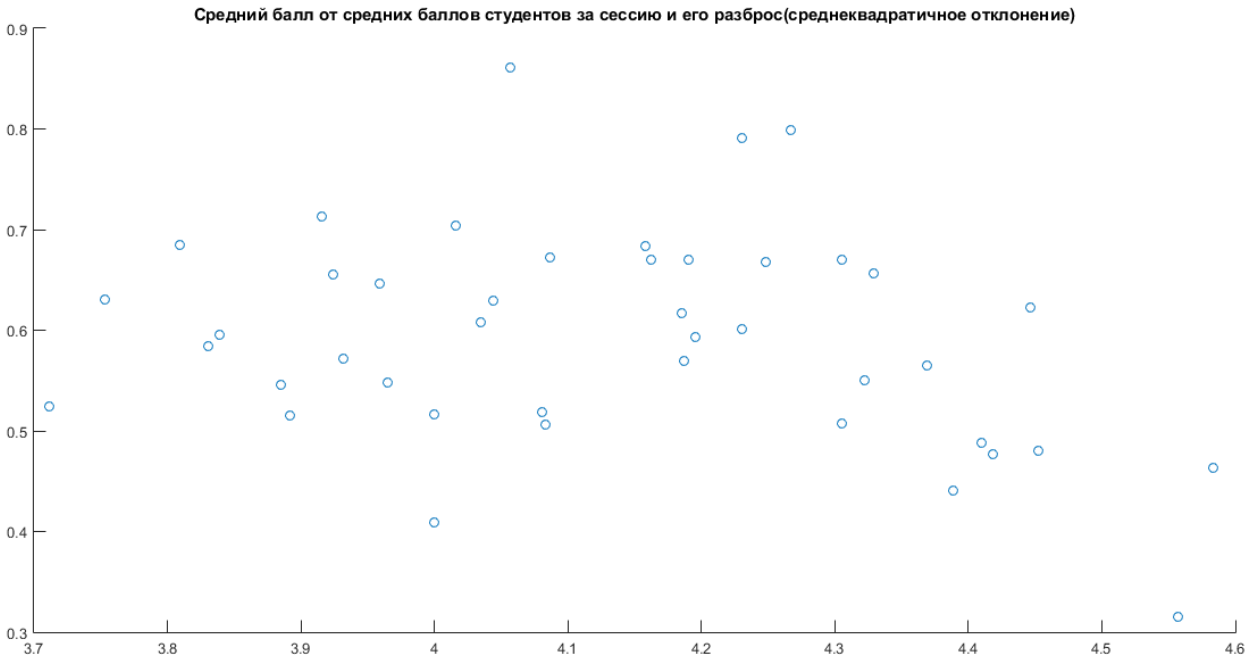
3.5 Средний балл и среднеквадратичное отклонение баллов

Рассмотрим некоторый экзамен в некоторую сессию. Рассмотрим аттестацию по дисциплине “Алгебра” в зимнюю сессию 2012-2013 года у студентов 2010 года поступления по направлению подготовки 010400 ”Прикладная математика и информатика”. Пусть, к примеру, получен набор оценок [5,4,5,4,3,4,5,3,3,5,4,4]. От этих оценок мы можем посчитать среднее значение и среднеквадратичное отклонение от среднего значения. Именно эти 2 параметра и изображены в качестве кружка или крестика на графике. Также в сессию мы можем получить такой же набор, но не баллов за экзамен, а среднего балла студента в эту сессию. Аналогично, точка на графике - соответствующие параметры.

Теперь построим 3 графика. График 1 отражает наши параметры для всех экзаменов и средних баллов. Рисунок 2- гистограмма средних оценок из графика 1. График 3- это взятые отдельно, для наглядности, средние оценки от средних оценок студента за сессию и разброс средних оценок в сессию.







Исходя из вида графика 1, можем предположить, что статистическая зависимость между средним баллом и дисперсией носит параболический характер. Прослеживаемая тенденция качественно говорит о следующем.

Чем ближе средний балл к минимальному или максимальному, тем меньше становиться дисперсия. В общем, эта тенденция вполне обоснована ограниченным выбором оценок. Также данная тенденция схожа для разных сессий, студентов разных годов поступления и других различных факторах. Дополнительно следует упомянуть, что похожая зависимость была найдена в работе Сосницкого, Потанина и Шевелевой "Проблемы статистического анализа средней успеваемости студентов" [7].

Нижеследующая за графиком гистограмма (рис.2) среднеквадратичных отклонений и средних значений показывает, что при рассмотрении графика можно выделить 3 группы оценок успеваемости: до 3.6 баллов, затем от 3.6 до 4.6 и от 4.6 и более. Это значит, что есть группа общей успеваемости (от 3.6 до 4.6), ”тройки” (до 3.6 балла) и “отлично” (более 4.6 балла). Группа общей успеваемости наибольшая по количеству, характеризуется линейно небольшим наклоном. Её коэффициент корреляции r=-0.31, а данные значимо коррелируют. “Тройки”- чем меньше средний балл, тем меньше разброс, т.е. тем больше троек. Значимо коррелируют данные, а линейно характеризуется коэффициентом r=0,56. “Отлично”- данные этой группы также значимо коррелируют с коэффициентом r=-0.94. Для этой группы характерно что, чем ближе средний балл к пятёрке, тем меньше дисперсия и, как следствие, больше пятёрок.

Важно заметить, что выше рассматривался балл за экзамен и разброс этих баллов в отличие, от средних баллов студента за сессию и разброса сессионных оценок. Для наглядности продемонстрируем это на простом примере. Пусть были выставлены за экзамен 3 пятёрки, 8 оценок хорошо и 5 троек. Соответственно имеем средний балл- 3.875, а среднеквадратичное

отклонение - 0.69. Это будет наша точка на графике. От этого стоит отличать средний балл студента за сессию и разброс этих баллов, что не относится к теме данной главы. Например, Сидоров получил{4,5,4,5}, Иванов-{3,3,5,5}. Соответственно имеем средний балл Сидорова 4.5 и среднеквадратичное отклонение 0.5, а у Иванова соответственно 4 и 1. Средний балл от средних баллов Иванова и Сидорова будет 4,25. Эти описания естественно взаимосвязаны, но первое делает акцент на успеваемость по экзамену, а второе - на успеваемость студентов.

Как видно из первого графика (рис.1), включение среднего балла (зелёные кружки) даёт нам “размытие” параболической зависимости. Рассмотрим наши зависимости с математической точки зрения. Для сравнения регрессионных моделей, будем использовать скорректированный коэффициент детерминации R^2-adjusted. Специфика его использования в сравнениях моделей с разным числом факторов. Как известно, по шкале Чеддока считается, что для приемлемых моделей R^2 более 50%, более 70% свидетельствует о высокой корреляции. А R^2порядка 90%-99% говорит о значительной взаимосвязи.

Рассмотрим интерполяцию всех средних баллов и дисперсий (красная парабола и красные крестики, рис.1). Объём этой выборки 229 значений. Эти данные значимо коррелируют с 99,5% уровнем. Линейная аппроксимация даёт коэффициент детерминации 21% (R^2 значим с 99% уровнем по F-тесту, $F=60,34 > F_{критического}=6,74$) . Параболическая- 41% (R^2 значим с 99% уровнем по F-тесту, $F=78,52 > F_{критического}=4,7$). Интерполяция полиномами большей степени даёт незначительный~ 1%-2% прирост в качестве модели. Степень

полинома усложняет модель, поэтому для баланса простоты и надёжности модели остановимся на параболической интерполяции.

$$y = -0.23\,x + 1.69 \qquad 21.05\%$$

$$y = -0.54\,x^2 + 4.29\,x - 7.68 \qquad 41.52\%$$

$$y = -0.32\,x^3 + 3.53\,x^2 - 12.74\,x + 15.92 \qquad 41.86\%$$

Рассмотрим взаимосвязь среднего балла от средних баллов студентов за сессию и их, средних баллов, разброс. Рисунок 3 или данные, выделенные зелёными кружками на графике 1. Коэффициент корреляции= -0.26 и, с уровнем значимости в 95%, мы отвергаем гипотезу о ненулевой корреляции. Т.е. наши данные не взаимосвязаны. Как следствие, нет необходимости в построении регрессионной модели зависимости. Из этого мы можем сделать вывод, что разброс средних баллов студентов за сессию не связан с общим уровнем сдачи сессии.

Рассмотрим теперь взаимосвязь среднего балла за экзамен и разброс экзаменационных оценок. Объём этой выборки 187 значений. Данные коррелируют, причём с 99.5% уровнем значимости. Для линейной модели имеем R^2 = 24% (R^2 значим с 99% уровнем по F-тесту, F=58,42> $F_{критического}$=6,77), для параболической скорректированный коэффициент детерминации составляет уже 57% (R^2 значим с 99% уровнем по F-тесту, F=121,95> $F_{критического}$=4,72). При увеличении степени полинома происходит незначительная прибавка в качестве. Прибавка порядка 1%-2%. Поэтому, аналогично вышеописанному рассуждению, остановимся на параболической модели. Причём в данном случае мы можем говорить об адекватном описании взаимосвязи данных нашей параболической моделью.

$$y = -0.24\,x\ + 1.74 \qquad 24.61\%$$

$$y = -0.64\,x^2 + 5.15\,x - 9.43 \qquad 56.67\%$$

$$y = -0.25\,x^3\ + 2.53\,x^2\ - 8.11\,x\ + 8.96 \qquad 57.44\%$$

Нижеследующий график отражает зависимость среднего балла студента за сессию и разброса этих баллов. График построен по данным всех сессий и всех студентов. Как видно из графика, мы не можем говорить о параболической, как в вышеописанном случае, зависимости. Скорее, этот график свидетельствует о независимости данных.

Вывод:

1. Успеваемость, по экзаменам, разбивается на три категории, по среднему баллу. До 3.6 балла, от 3.6 до 4.6 и со средним баллом более 4.6

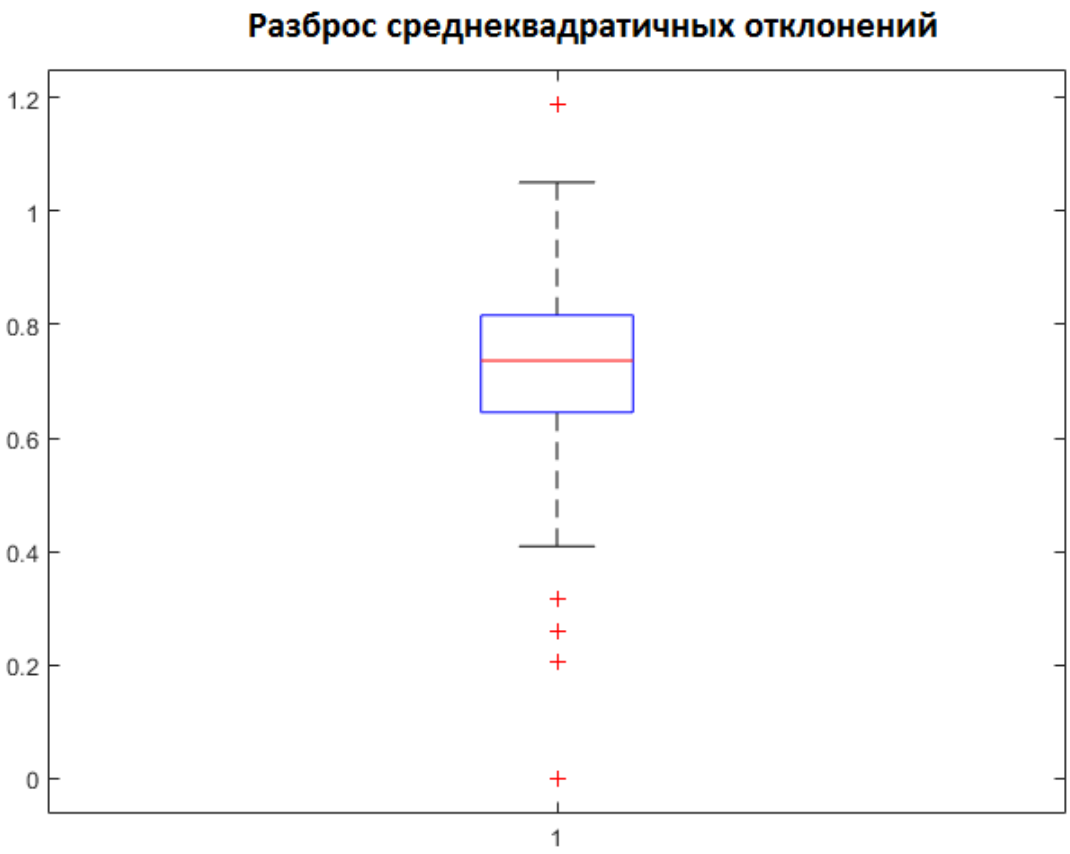
2. Разброс средних баллов студентов за сессию не связан со средним баллом от всех средних баллов за сессию, т.е. с общим уровнем сдачи сессии
3. Средний балл за предмет и разброс экзаменационных оценок значимо коррелируют и адекватно описываются параболической регрессионной моделью:

$$y = -0.64\,x^2 + 5.15\,x - 9.43 \qquad R^2 = 56.67\%$$

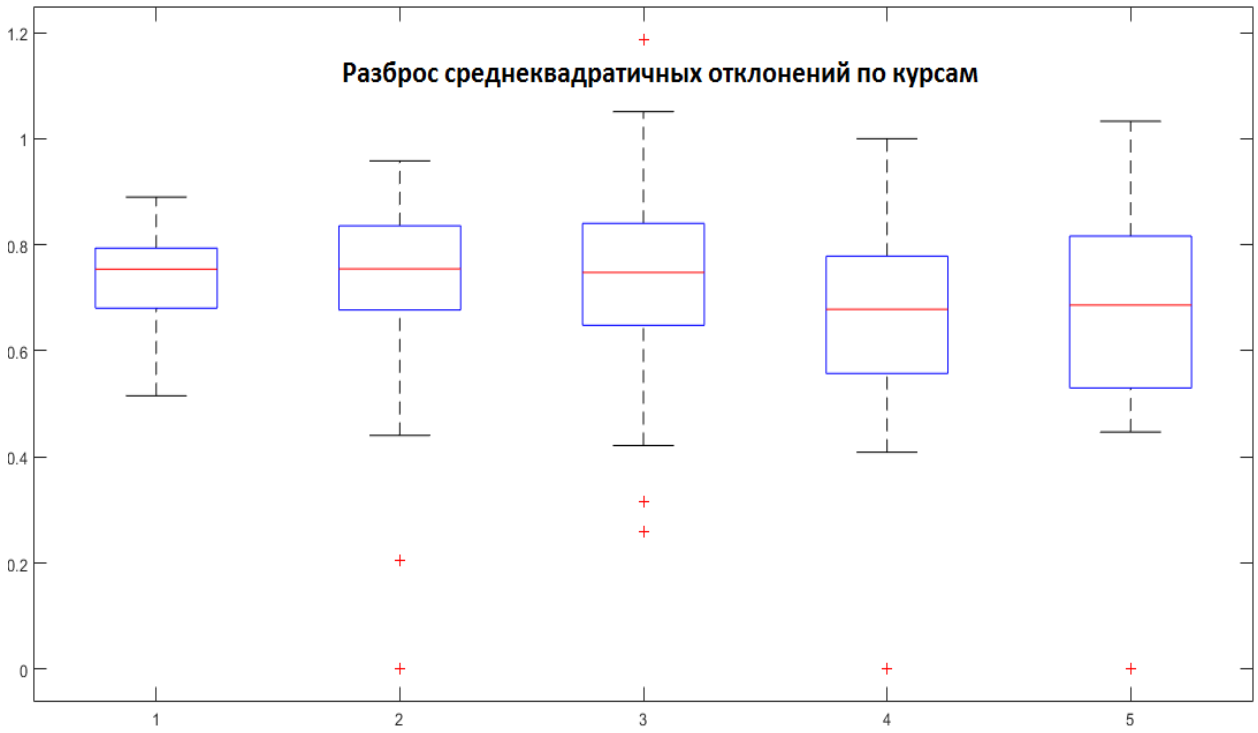
3.6 Исследование среднеквадратичных отклонений отметок за экзамен

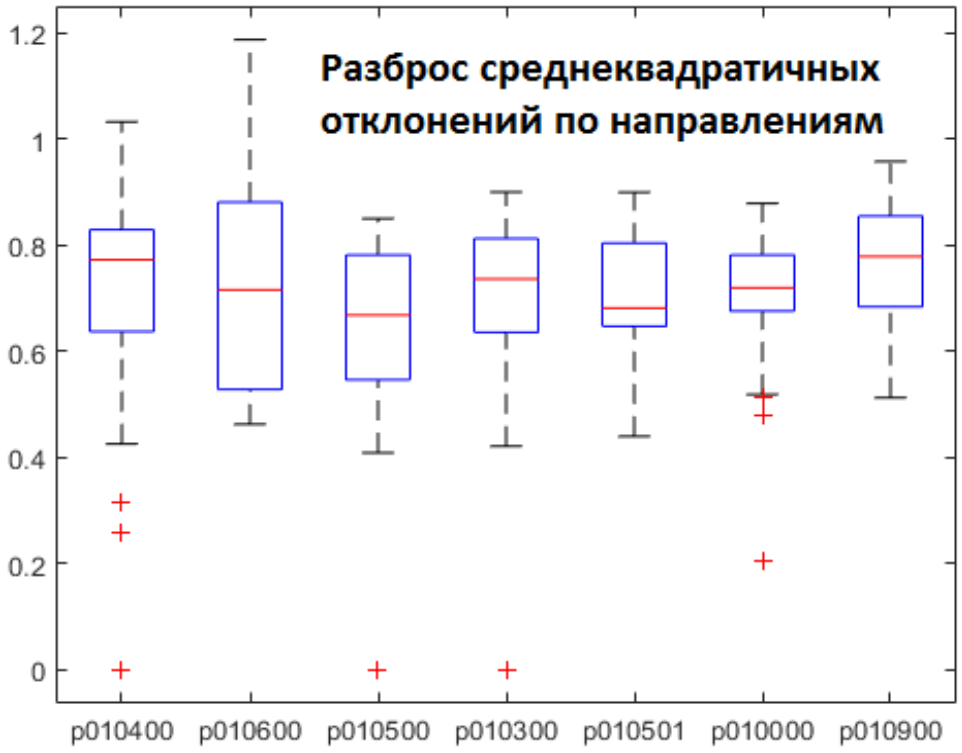
Предыдущий параграф показал взаимосвязь среднего балла и среднеквадратичных отклонений, посчитанных для выборок оценок экзаменов. Этот параграф рассмотрит общую картину распределения среднеквадратичных отклонений, посчитанных для выборки из оценок за экзамен и выборок среднего сессионного балла студента.

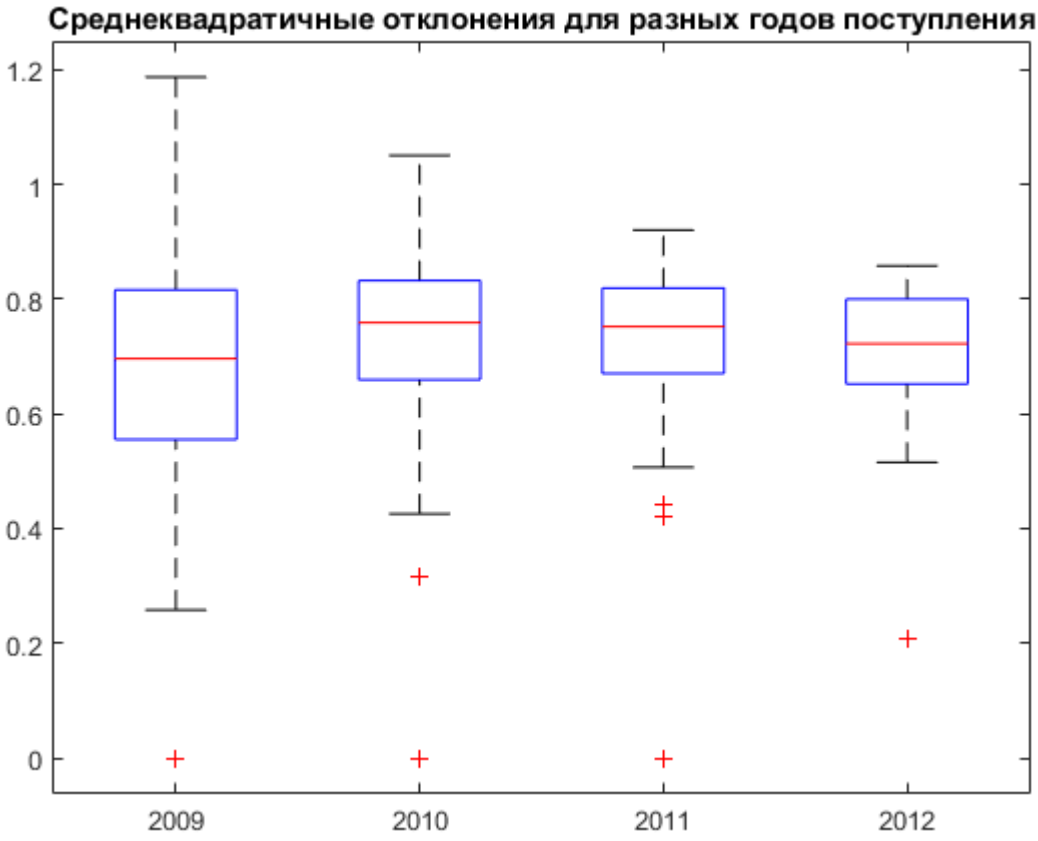
Нижеследующие графики представляют общую статистику поведения среднеквадратичного отклонения. Графики представляют собой “ящики с усами”, построенные с помощью функции MATLAB boxplot.Исследуются среднеквадратичные отклонения с сечением по курсу обучения, направлению подготовки и году поступления.


Разброс среднеквадратичных отклонений
1.2
1
0.8
0.6
0.4
0.2
0
1


Гистограмма среднеквадратичных отклонений
50
45
40
35
30
25
20
15
10
5
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4


Разброс среднеквадратичных отклонений по курсам
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3
4
5


Разброс среднеквадратичных отклонений по направлениям
1.2
1
0.8
0.6
0.4
0.2
0
p010400
p010600
p010500
p010300
p010501
p010000
p010900



Как видно из гистограммы, около половины всех значений лежит в пределе от 0.6 до 0.8. Причём, похожая картина сохраняется для разных курсов обучения, годов поступления и направлений обучения. Качественно это означает, что общая картина даёт разброс +-0.7 балла.

Гистограмма отклонений говорит о возможном согласии данных с нормальным распределением. Тест Ярки-Бера призван осветить этот вопрос. Он состоит в проверке согласия выборки значений нормальному закону с неопределёнными параметрами. Этот тест реализован в системе MATLAB-jbtest. Данные возникают из общей сводной таблицы DescrStat. С 99% уровнем значимости распределение среднеквадратичных отклонений не противоречит нормальному распределению. Параметры нормального распределения рассчитываются точечно и доверительным интервалом с 95% уровнем значимости. Для нахождения значений используется функция

normfit.В ней применяется метод моментов для поиска параметров распределения. Доверительные интервалы строятся использованием соответствующих статистик и проверки на попадание в критическую область распределения Стьюдента или хи-квадрат, для соответствующих параметров. Ниже приводятся результаты вычислений:

$$P\left(s{\sim}N(\mu;\sigma^2)\right)=99.9\%$$

$$s{\sim}N(0.713;0.1656^2)$$

$$P(\mu\in\langle 0.6915|0.7346\rangle)=95\%$$

$$P(\sigma\in\langle 0.1517\,|0.1824\rangle)=95\%$$

Кроме общих закономерностей, графики также предоставляют нам данные о выбросах. Что бы выдвигать гипотезы и делать выводы данных недостаточно, но исключения вполне можно описать. Преимущество зимних экзаменов, вероятно, следует объяснить наличием большого числа данных по зимним сессиям.

Величина выборки	Год поступления	Учебный год	Сессия	Направление	Курс	Предмет	Средний балл	Разброс
146	2010	2011х2012	Зима	010400	2	Курс по выбору	5	0
42	2011	2012х2013	Зима	010300	2	Курс по выбору	5	0

23	2009	2012x2013	Зима	010500	4	Курс по выбору	4	0
286	2012	2013x2014	Зима	010000	2	Компьютерные сети	4,95	0,20
15	2009	2011x2012	Зима	010400	3	Технологии сети Интернет	4,93	0,25
14	2010	2012x2013	Лето	010400	3	Средний балл зачётки	4,55	0,31
13	2009	2011x2012	Зима	010600	3	Дифференциальные и интегральные уравнения	4,07	1,18

- Среди выбросов нет ни одного с первого курса
- 3 предмета с нулевой дисперсией, двумя оценками отлично и одной хорошо - курсы по выбору
- 2 экзамена с незначительным среднеквадратичным отклонением и отличными оценками относятся к информационным технологиям
- Одно малое среднеквадратическое отклонение со средним баллом 4.5 характеризует группу студентов, успешно сдавшую сессию
- На сдаче дифференциальных и интегральных уравнений был значительный разброс оценок
- Зависимостей от направления, года поступления, учебного года и, что немаловажно, величины выборки не прослеживается

Вывод:

- Среднеквадратичные отклонения ведут себя схожим образом для сечений разными курсами обучения, годами поступления и направлениями обучения
- Среди выбросов из общей картины среднеквадратических отклонений нет ни одного с первого курса
- Среднеквадратичные отклонения оценок за экзамен описываются нормальным распределением с параметрами:

$$s \sim N(0.713; 0.1656^2)$$

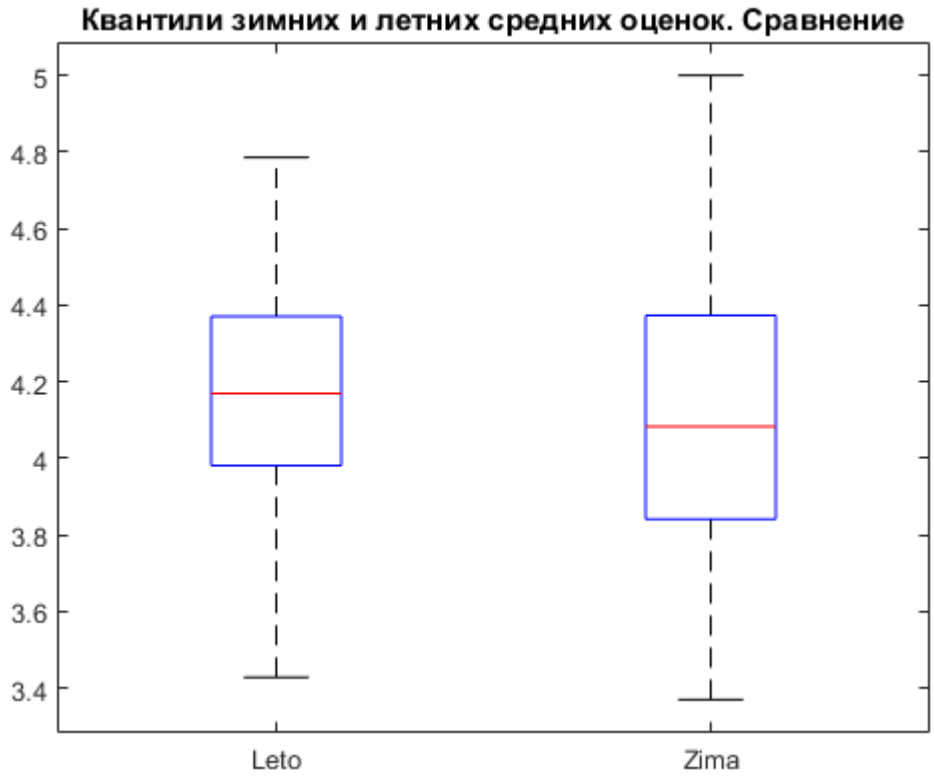
$$P(\mu \in \langle 0.6915 | 0.7346 \rangle) = 95\%$$

$$P(\sigma \in \langle 0.1517 | 0.1824 \rangle) = 95\%$$

3.7 Сравнение успеваемости зимних и летних сессий

Рассмотрим экзамен и оценки за него. Это будет некоторая выборка баллов. От неё мы можем взять средний балл. Теперь рассмотрим такие средние баллы за все экзамены из наших данных, с разделением по фактору времени года. Т.е. рассмотрим средний балл за экзамены зимних и летних сессий. Воспользуемся функцией boxplot в системе MATLAB. Получили график 0%, 25%, 50%,75%,100%перцентилей. Особенно нас интересуют 25% 50% и 75% перцентили, интересуют для исследования сходств и различий в общем уровне сдачи экзаменов. Для летних сессий они 4, 4.18 и 4.36 соответственно. Для зимних- 3.84, 4.03, 4.37. Видим, что 75% квантили практически совпадают. Разница 50% составляет 0.15 балла, а 25% - 0.16. Получается, что летом баллы за экзамен в среднем выше. Из этого можно поставить закономерный вопрос - однородны ли эти выборки вообще?

Перефразируя, сдаются ли экзамены летом и зимой по схожим статистическим законам?



Рассмотрим этот вопрос с математической точки зрения. Воспользуемся критерием однородности двух выборок.

По Критерию Колмогорова-Смирнова, точнее его реализации в прикладном пакете MATLAB kstest2, с 95% уровнем значимости мы отвергаем нулевую гипотезу о том, что выборки зимних и летних сессий принадлежат к одному закону распределения. Уровень значимости, при котором нашу гипотезу о том, что выборки распределены по одному закону распределения, мы бы не отвергли, составляет 66.5%.

Обозначим выборку летних оценок за x, зимних - за y. Проведём следующее преобразование, и сравним снова. Данные могут принадлежать одному закону распределения, но с разными параметрами. Проведём стандартизацию данных. Если мы не отвергнем гипотезу однородности, то в дальнейшем можем исследовать принадлежность к некоторому общему закону распределения с неизвестными параметрами.

$$x = \frac{x - M[x]}{\sqrt{D[x]}}$$

$$y = \frac{y - M[y]}{\sqrt{D[y]}}$$

Тест Колмогорова-Смирнова отверг однородность оценок с 95% уровнем значимости.

Возникает вопрос о том, подчиняются ли отметки одной из групп, или всей выборки, нормальному распределению. Тест Ярки-Бера состоит в согласии выборки значений случайной величины нормальному закону распределения. Воспользуемся его реализацией в пакете MATLAB- jbtest. С 95% уровнем значимости средний балл от отметок за экзамен летних, зимних и всей выборки противоречит нормальному распределению. Следует отметить специфику теста Ярки-Бера. Этот тест не следует использовать на малых выборках. Для проверки гипотезы о соответствии выборки нормальному закону распределения на малых выборках воспользуемся использовать функцией lillietest.

Тест Лиллиефорса заключается в проверке гипотезы о не противоречии генеральной выборки нормальному закону распределения с неизвестными параметрами. Воспользуемся его реализацией в программе MATLAB и

получим, что с 95% уровнем значимости наши выборки противоречат нормальному распределению. Из чего мы можем заключить, что распределение среднего балла от оценок за экзамен не согласуется с нормальным распределением.

Выводы:

1. Общий уровень сдачи летних экзаменов выше зимних
2. Распределение среднего балла от оценок за экзамены летних, зимних и общей выборки сессий не согласуется с нормальным распределением
3. Средние баллы успеваемости в летние и зимние экзамены неоднородны

Выводы

На основании проведённых исследований можно привести нижеследующие выводы.

1. Значимость различных факторов, влияющих на успеваемость

Уровень значимости	**Фактор**
Незначимый	• целевой набор
Условно-значимый	• регион поступления • направление обучения • фактор поступления по олимпиаде (100 баллов по ЕГЭ) • особые обстоятельства зачисления
Значимый	• форма обучения (бюджет/коммерция) • зачисление по олимпиаде без экзаменов
Высокий уровень значимости	• проходной балл ЕГЭ • форма обучения
Особо значимый фактор	• баллы ЕГЭ

2. Данные успеваемости говорят о том, что отметками единого государственного экзамена можно объяснить вариацию успеваемости. Следует отметить, что объяснимость успеваемости отметками за единый государственный экзамен с курсом обучения понижается

3. Факторы, влияющие на успеваемость студентов по информационным технологиям и математическим дисциплинам схожи, в отличие от математических дисциплин или предметов информационных технологий и гуманитарных дисциплин. Курсы обязательные и необязательные к изучению характеризуются однородностью выборок

4. Отметки успеваемости за экзамен, как и средние баллы студента за сессию согласуются с нормальным законом распределения

5. Средний балл за предмет и разброс (среднеквадратичное отклонение) экзаменационных оценок значимо коррелируют и адекватно описываются параболической регрессионной моделью. Для средних баллов студента за сессию схожей тенденции не наблюдается

6. Среднеквадратичные отклонения оценок за экзамен описываются нормальным распределением

7. Данные успеваемости говорят, что общий уровень сдачи летних экзаменов выше зимних. Также следует упомянуть, что средние баллы успеваемости в летние и зимние экзамены неоднородны

Заключение

В квалификационной работе была рассмотрена задача исследования успеваемости с помощью вероятностно-статистического аппарата. Особенностью исследования является рассмотрение успеваемости в контексте временной динамики.

В начале работы приводится введение в задачу исследования успеваемости, обзор литературы и фактическая постановка задачи. В первой главе формируется структура проведения исследования и описывается предметная область. Затем выделяются значимые сущности, из которых формируется информационно-логическая модель успеваемости. После чего, на основании модели, естественным образом возникают задачи, необходимые для исследования успеваемости. Кроме вышесказанного во второй главе приведён параграф, описывающий первичную обработку и организацию данных. Во второй главе приводятся математические методы, используемые в дальнейшем. *В* третьей главе, проводятся вероятно-статистические исследования решающие задачи, поставленные в предыдущей главе. На основе регрессионной модели проводится анализ факторов, влияющих на успеваемость. На основании чего факторы ранжируются. Выделяется особо значимый фактора – баллы ЕГЭ при поступлении. В следующем параграфе проводится более углублённое исследование связи единого экзамена и успеваемости. Далее был проведён анализ взаимосвязи предметов и их оценок. Нормальное распределение имеет фундаментальное значение в математической статистике, поэтому в очередном параграфе проверяется гипотеза о согласии оценок с нормальным распределением. Проверяется как распределение баллов за

экзамен, так и средняя оценка студента в сессию. Следующий параграф частично подтверждает результаты статьи [7] о значимой связи среднего балла за экзамен и среднеквадратического отклонения этих баллов, кроме этого проверяется связь среднего балла студента в сессию с соответствующим разбросом. Последующий параграф проясняет поведение среднеквадратичных отклонений отметок за экзамен в разрезе различных факторов, также проводится анализ выбросов из общей картины. К примеру, практически нулевые дисперсии отметок за экзамен. Заключительный параграф говорит о сравнении успеваемости в зимние и летние сессии. Проводится анализ их однородности. Проведённые исследования позволяют, как более глубоко изучить академическую успеваемость, так и послужить основой для принятия решений.

Дальнейшее развитие работы может осуществляться как в проведении подобных исследований по данным другого контингента универсантов, например, студентов других факультетов или других ВУЗов, так и в расширенных постановках задач.

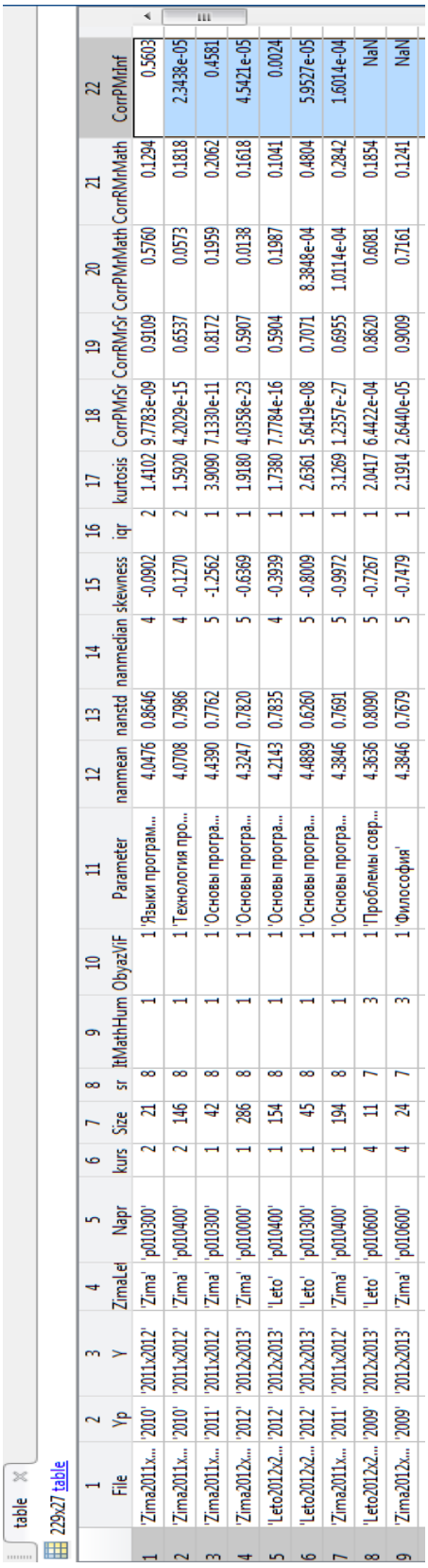
Список литературы

1. Богомолов А. И., Деркаченко В. Н., Арюткина Т. А. Прогнозирование успеваемости обучающихся по специальным дисциплинам на основе регрессионных уравнений // Известия высших учебных заведений. Поволжский регион. №1(9) 2009, с. 124-132.
2. Бодряков В. Ю., Торопов А. П., Фомина Н. Г. Анализ успеваемости как прогноз успешной деятельности выпускников математического факультета педагогического университета // Педагогическое образование в России. № 2 / 2010, с. 130-140
3. Герасименко П. В., Руслан С.В. Исследование динамики изменения успеваемости по математическим дисциплинам студентов экономических специальностей ПГУПС // Известия ПГУПС. №1(34)/2013, с. 215-221
4. Сосницкий В.Н., Потанин Н.И Вероятностный подход к анализу успеваемости студентов // Фундаментальные исследования. № 8-3 / 2014, с. 734-738
5. Хавенсон Т. Е., А. А. Соловьева Связь результатов Единого государственного экзамена и успеваемости в вузе // Вопросы образования. № 1 / 2014, с. 176-199
6. Грязева Е.Д., И.Б. Губанцева, Э.М. Попов Психофизиологические характеристики и успеваемость студентов первокурсников не физкультурного ВУЗа // Известия Тульского государственного университета. Физическая культура. Спорт. № 1 / 2014, с. 45-51

7. Сосницкий В.Н., Потанин Н.И., Шевелева Л.В. Проблемы статистического анализа средней успеваемости студентов // Фундаментальные исследования.№ 10-2 / 2013, с.316-320
8. Пермякова Т. М., М. С. Шевелева ЕГЭ как предиктор успешности изучения английского языка в вузе // Вестник Ленинградского государственного университета им. А.С. Пушкина. № 2 / том 1 / 2013, с.205-214
9. Бодряков В. Ю., Е. Н. Нигматуллина, Н. Г. Фомина Исследование структуры интеллекта студентов-математиков: проблематика успешного обучения // Педагогическое образование в России. № 2 / 2010, с.36-42
10. Пермякова Т. М., М. С. Шевелева Лингводидактика и методика обучения языку // Вестник Ленинградского государственного университета имени А.С. Пушкина. 2013 №02, с.205-214
11. Трифонов А.Ю., А.А. Михальчук Сравнительный статистический анализ оценки математических знаний студентов первого курса // Известия Томского политехнического университета. № 5 / том 308 / 2005, с.212-216
12. Степанов В.М., О.Ю. Трошин, Е.Л. Тихонова, М.Ф. Чурбанов Статистические характеристики перехода «абитуриент – студент» на химическом факультете ННГУ // Вестник Нижегородского университета им. Н.И. Лобачевского.№ 3-1 / 2012, с.11-16

Приложения

Пример сводной таблицы DescrStat



table

229x27 table

	1 File	2 Yp	3 Y	4 ZimaLet	5 Napr	6 kurs	7 Size	8 sr	9 ItMathHum	10 ObyazVF	11 Parameter	12 nanmean	13 nanstd	14 nanmedian	15 skewness	16 iqr	17 kurtosis	18 CorrPMSr	19 CorrRMSr	20 CorrPMrMath	21 CorrRMrMath	22 CorrPMrInf
1	'Zima2011x...	'2010'	'2011x2012'	'Zima'	'p010300'	2	21	8	1	1	'Языки програм...	4.0476	0.8646	4	-0.0902	2	1.4102	9.7783e-09	0.9109	0.5760	0.1294	0.5603
2	'Zima2011x...	'2010'	'2011x2012'	'Zima'	'p010400'	2	146	8	1	1	'Технологии про...	4.0708	0.7986	4	-0.1270	2	1.5920	4.2029e-15	0.6537	0.0573	0.1818	2.3438e-05
3	'Zima2011x...	'2011'	'2011x2012'	'Zima'	'p010300'	1	42	8	1	1	'Основы програ...	4.4390	0.7762	5	-1.2562	1	3.9090	7.1330e-11	0.8172	0.1959	0.2062	0.4581
4	'Zima2012x...	'2012'	'2012x2013'	'Zima'	'p010000'	1	286	8	1	1	'Основы програ...	4.3247	0.7820	5	-0.6369	1	1.9180	4.0358e-23	0.5907	0.0138	0.1618	4.5421e-05
5	'Leto2012x2...	'2012'	'2012x2013'	'Leto'	'p010400'	1	154	8	1	1	'Основы програ...	4.2143	0.7835	4	-0.3939	1	1.7380	7.7784e-16	0.5904	0.1987	0.1041	0.0024
6	'Leto2012x2...	'2012'	'2012x2013'	'Leto'	'p010300'	1	45	8	1	1	'Основы програ...	4.4889	0.6260	5	-0.8009	1	2.6361	5.6419e-08	0.7071	8.3848e-04	0.4804	5.9527e-05
7	'Zima2011x...	'2011'	'2011x2012'	'Zima'	'p010400'	1	194	8	1	1	'Основы програ...	4.3846	0.7691	5	-0.9972	1	3.1269	1.2357e-27	0.6955	1.0114e-04	0.2842	1.6014e-04
8	'Leto2012x2...	'2009'	'2012x2013'	'Leto'	'p010600'	4	11	7	3	1	'Проблемы совр...	4.3636	0.8090	5	-0.7267	1	2.0417	6.4422e-04	0.8620	0.6081	0.1854	NaN
9	'Zima2012x...	'2009'	'2012x2013'	'Zima'	'p010600'	4	24	7	3	1	'Философия'	4.3846	0.7679	5	-0.7479	1	2.1914	2.6440e-05	0.9009	0.7161	0.1241	NaN

Пример таблицы начальных данных

оссийской Федерации или страны (согласно Приложению №1) по адресу постоянной регистрации (гражданства)		
ип населенного пункта (согласно Приложению №3) по месту постоянной регистрации (гражданства)		
населенном пункте по месту постоянной регистрации **(заполняется при наличии соответствующих данных)**		
Тип оплаты обучения (Бюджет - 1, Внебюджет - 0)		
и геометрия	Экзамен 1	Оценки за 1-ю сессию
ческий анализ I	Экзамен 2	
ограммирования	Экзамен 3	
ие экзамена	Экзамен 4	
ие экзамена	Экзамен 5	
ие экзамена	Экзамен 6	
ие экзамена	Экзамен 7	
л по ЕГЭ	Математика	Баллы по предметам, учитывавшимся при зачислении
в аттестате		
л по ЕГЭ	Информатика и ИКТ	
в аттестате		
л по ЕГЭ	Физика	
в аттестате		
л по ЕГЭ	Химия	
в аттестате		
л по ЕГЭ	География	
в аттестате		
л по ЕГЭ	Биология	
в аттестате		
л по ЕГЭ	Русский язык	
в аттестате		
л по ЕГЭ	Обществознание	
в аттестате		
л по ЕГЭ	Иностранный язык	
в аттестате		